M321 5, 6 and 7

THE OPEN UNIVERSITY

Mathematics: A Third Level Course

Partial Differential Equations of Applied Mathematics Units 5, 6 and 7

Initial Value Problems
Fourier Series
Overhead Wires



THE OPEN UNIVERSITY

Mathematics: A Third Level Course

Partial Differential Equations of Applied Mathematics
Units 5, 6 and 7

INITIAL VALUE PROBLEMS
FOURIER SERIES
OVERHEAD WIRES

Prepared by the Course Team

The Open University Press

The Open University Press, Walton Hall, Milton Keynes.

First published 1974.

Produced in Great Britain by
Technical Filmsetters Europe Limited, 76 Great Bridgewater Street, Manchester M1 5JY.

ISBN 0 335 01251 5.

This text forms part of the correspondence element of an Open University Third Level Course. The complete list of units in the course is given at the end of this text.

For general availability of supporting material referred to in this text, please write to the Director of Marketing, The Open University, P.O. Box 81, Milton Keynes, MK7 6AT.

Further information on Open University courses may be obtained from The Admissions Office, The Open University, P.O. Box 48, Milton Keynes, MK7 6AB.

1.1

Unit 5 Finite-Difference Methods I: Initial Value Problems

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differential Equations* (Oxford, 1971).

H. F. Weinberger, *A First Course in Partial Differential Equations* (Blaisdell, 1965).

It is essential to have these books: the course is based on them and will not make sense without them. They are referred to in the text as S and W respectively.

Unit 5 is based on S: Chapter 1, pages 6 to 8 and Chapter 2, pages 10 to 24, 32 to 40 and 46 to 52.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics*. References to Open University courses in mathematics take the form:

Unit M100 13, Integration II for the Mathematics Foundation Course,
Unit M201 23, The Wave Equation for the Linear Mathematics Course.

5.0 INTRODUCTION

The various analytical methods of obtaining solutions to partial differential equations, some of which are discussed in this course, can be applied only to a restricted range of problems. For example, the initial value problem

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) + (\sin xt)u(x,t) = 0 \qquad 0 < x < 1, t > 0$$

$$u(0,t) = u(1,t) = 0 \qquad t \geqslant 0$$

$$u(x,0) = f(x) \qquad 0 \leqslant x \leqslant 1,$$

cannot be solved by the method of separation of variables or by means of integral transforms (e.g. Laplace and Fourier transforms). Even when solutions can be found analytically, difficulties often arise in evaluating them numerically. For instance, the solution may involve finding the Fourier coefficients of the function which appears in an initial condition (*Unit M201 32, The Heat Conduction Equation*, Section 32.2). Each coefficient involves an integral which implies that, if we are aiming for full numerical precision, an infinitude of integrals must be evaluated. Unless the given function is particularly simple (for example, a polynomial, exponential, or trigonometric function) only a finite number of these integrations can be carried out in a finite time. We are therefore reduced to approximating the Fourier series (e.g. by truncating it), and might even have to estimate approximately each of the separate integrals in the series. As a result we would end with only an approximate solution to our original problem.

In this course we present some numerical methods which can be applied to a wide class of problems, but which inherently give only approximate values for the solutions. We do not claim that the numerical methods we shall illustrate can be applied to all problems involving partial differential equations or that they are the only methods available, but they are currently the methods most frequently used and most widely applicable.

In *Unit M201 7, Recurrence Relations* and *Unit M201 8, Numerical Solution of Simultaneous Algebraic Equations* we saw that, in numerical mathematics, attention had to be paid on the one hand to the formulation of the problem, and on the other hand to the method of solution.

We have to consider the possibility of *inherent instability* in the problem, in which case small changes in the data make large changes in the solution. We have met this previously in this course, and we described such a problem as not being *continuous with respect to its data.*

Throughout this unit we shall concentrate on *initial value problems*, i.e. problems in which the value of the solution (or its normal derivative) is given along the *initial line* ($t = 0$). Initial value problems in which a boundary condition is specified for $t > 0$ at each point of the spatial boundary are properly posed for parabolic equations. In particular they exhibit continuity, and so do not suffer from inherent instability, as we have seen in *Unit 3, Elliptic and Parabolic Equations*. Similarly, the Cauchy problem (i.e. the problem as above, but with the solution and its normal derivative *both* specified on the initial line) is properly posed for hyperbolic equations. (We usually refer to these problems as *initial-boundary value problems*.) By analogy with the treatment of initial-value problems for ordinary differential equations, discussed in *Unit M201 21, Numerical Solution of Differential Equations* we use approximating recurrence relations in a step-by-step process leading to a simple system of simultaneous algebraic equations. Each of the unknowns in the system of equations represents an approximate value of the solution to the differential equation at some predetermined point. We might expect that as the number of points is increased the numerical solution becomes more accurate. We shall employ methods for which we can prove (under suitable conditions) that this is the case.

The recurrence relations are obtained by *finite-difference methods*, and we shall discuss *explicit* and *implicit* schemes. In the former, as the name implies, each equation

expresses one unknown value in terms of known values. In the latter, however, several unknown values must be solved for simultaneously.

Following the discussion of the numerical solution of ordinary differential equations in *Unit M201 21*, we might expect that some of our methods exhibit *induced instability*. In this unit we shall illustrate such instability with specific numerical examples, and as a result we shall gain some insight into the factors which induce this phenomenon. We shall support this insight by a full theoretical treatment in *Unit 8, Stability*.

In the main part of this unit we shall produce and use only the simplest finite-difference equations for parabolic and hyperbolic equations and, if necessary, for the initial and boundary conditions. More accurate schemes can be constructed, but we shall not discuss them here. Their existence illustrates how the numerical analyst is constantly searching for methods of greater efficiency and wider applicability.

Numerical analysis has grown rapidly over the past few decades owing to the increasing availability of high-speed digital computers. The methods we shall investigate are designed primarily for such machines and in some SAQs in this unit you will be asked to run simple computer programs. These programs have already been written and are available as library routines; and the investigation of the results of computer runs using these programs is an integral part of the course. Occasionally, you may also be asked to *write* small programs in BASIC, but we shall not make it mandatory for you to do so.

5.1 ELEMENTARY FINITE-DIFFERENCE APPROXIMATIONS

We begin our work in numerical analysis by revising some ideas which form the basis of the finite-difference method as applied to partial differential equations. We shall need to use Taylor series (*Unit M100 14, Sequences and Limits II* and *Unit M201 14, Bilinear and Quadratic Forms*) and finite differences (*Unit M100 4, Finite Differences*).

To see how these ideas are used in practice we give a broad outline of the finite-difference method before going into details. The method requires us, at first, to choose a finite set of values for each of the domain variables in the problem. We usually choose these values in a systematic way in order to simplify the resulting theory and practice. For example, if we consider a problem with just two independent variables, denoted by x and t, then we take the sets of values $\{x_i : i = 0, 1, \ldots\}$ and $\{t_j : j = 0, 1, \ldots\}$, where successive members of each set are related by the formulas

$$x_{i+1} = x_i + h$$

and

$$t_{j+1} = t_j + k.$$

Here h and k are constants which we can choose at our discretion, subject to certain conditions, as we shall see. x_0 and t_0 are chosen so that the set of points

$$\{(x_i, t_j)\} = \{(x_0 + ih, t_0 + jk)\}$$

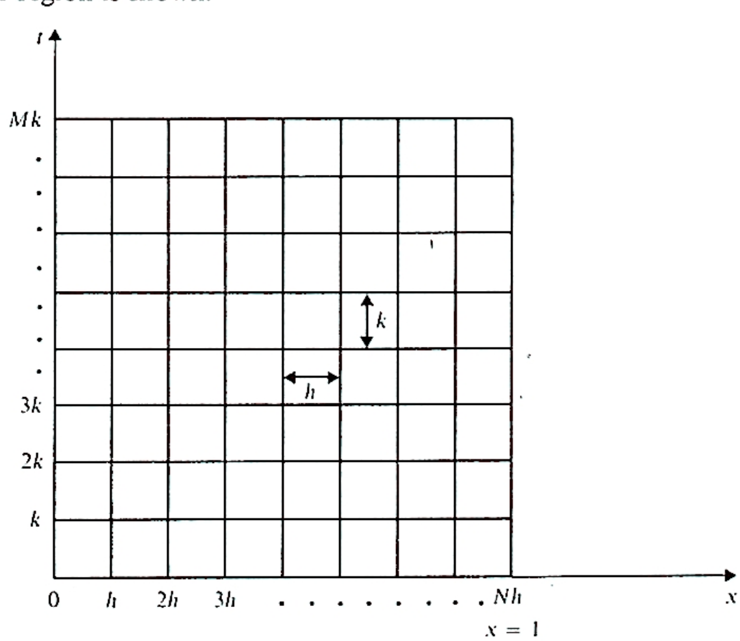
"covers" a suitable domain. In practice, we frequently choose coordinates such that $(x_0, t_0) = (0, 0)$. We shall sometimes denote the point (x_i, t_j) simply as i, j where x_0, t_0, h and k are understood. We refer to the set of points $\{(x_i, t_j)\}$ as a **mesh** or **grid**, and we call h and k the **mesh lengths** in the x- and t-directions respectively. The values $u(x_i, t_j)$ of a function u at the *mesh points* are called **pivotal values.**

Since it is possible to perform only a finite number of calculations in a finite time (even by computer!) we have to restrict the solution domain for any initial value problem that we care to tackle. From now on we shall assume that the mesh is finite; say $\{(x_i, t_j) : i = 0, 1, 2, \ldots, N \text{ and } j = 0, 1, 2, \ldots, M\}$.

It is always possible to measure x in suitable units so that $x_N = 1$. (This is discussed in *S: pages 9 and 10*, which is not, however, set as a reading passage.) Then

$$h = \frac{1}{N}$$

where $N + 1$ is the number of mesh points in the x-direction. The solution domain for the finite-difference calculations for two independent variables x and t in a rectangular region is shown.



The partial derivatives in the partial differential equation to be solved are approximated at each point by finite-difference formulas based on the chosen mesh and obtained using Taylor approximations. You have seen these ideas before, applied to ordinary differential equations in Section 21.2 of *Unit M201 21*.

Taylor's Theorem (*Unit M100 14*) is that

$$u(x + h) = u(x) + hu'(x) + \ldots + \frac{h^n}{n!} u^{(n)}(x) + C_n(h)$$

where

$$|C_n(h)| \leqslant \frac{1}{(n + 1)!} B_{n+1} |h^{n+1}|,$$

provided

$$|u^{(n+1)}(\bar{x})| \leqslant B_{n+1} \qquad \bar{x} \in [x, x + h]$$

and $u^{(n+1)}$ is continuous throughout $[x, x + h]$—or $[x + h, x]$ if $h < 0$.

This is quite a mouthful and we usually abbreviate it to

$$u(x + h) = u(x) + hu'(x) + \ldots + \frac{h^n}{n!} u^{(n)}(x) + O(h^{n+1}),$$

where $O(h^{n+1})$ is read as (*terms of*) *order* h^{n+1}. $O(h^{n+1})$ means loosely that h^{n+1} is the lowest power of h to appear in the expression it replaces. Note that O is *not* a function.

More precisely, we define the notation as follows. We say that

$$f(h) = O(g(h)) \quad \text{as} \quad h \rightsquigarrow 0,$$

if $\exists$ constants $\eta > 0$, $K > 0$ such that

$$|f(h)| \leqslant K|g(h)| \qquad h \in [-\eta, \eta].$$

For example, if

$$f(h) = 3h^2,$$

then, clearly we may say that $f(h) = O(h^2)$. In our case, we see (by Taylor's Theorem) that the remainder in the nth Taylor approximation is $O(h^{n+1})$ as $h \rightsquigarrow 0$ under the stated conditions, by choosing

$$K = \frac{B_{n+1}}{(n + 1)!}.$$

Taylor's Theorem is an important tool because it enables a function to be approximated in a given neighbourhood $[x - h, x + h]$ of a point x by a polynomial in h, and polynomials are relatively easy to handle. In practice we shall use Taylor's Theorem without specifying that u satisfies the conditions.

READ S: page 6, line 15 **Finite-difference approximations to derivatives** *to page 8, the end of Chapter 1.*

We have not asked you to read the first few pages of *S* dealing with elliptic equations (which are not discussed in this unit), but if you have time you might start reading from *S: page 4* **Parabolic and hyperbolic equations.**

Notes

(i) *S: page 6, Equations (1.3), (1.4) and (1.5)*

The expression represented by $O(h^4)$ is the difference between the true value of the left-hand side and the approximation. It is therefore called the *local truncation error* of the approximation since we have truncated the Taylor series to obtain it.

The notation

$$\left(\frac{d^2u}{dx^2}\right)_{x=x}$$

means

$$\frac{d^2u(x)}{dx^2}.$$

(ii) *S: page 6, line −7*

The term involving the lowest power of h in the error is the important one, because in practice h is much smaller than unity and we are interested in the behaviour as h approaches zero.

(iii) *S: page 8, Equations (1.8), (1.9) and (1.10)*

These equations are basic finite-difference approximations to the partial derivatives

$$\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial t^2} \text{ and } \frac{\partial u}{\partial t}$$

evaluated at the point P with coordinates $(x, t) = (ih, jk)$. (Note that the origin of our mesh of points is $(x_0, t_0) = (0, 0)$.) We shall make great use of these formulas throughout this unit in approximating the partial derivatives in a differential equation.

General Comment

For a function of one variable, $u: R \longrightarrow R$, we have seen how forward differences may be expressed in terms of the **forward-difference operator** Δ_h, whose effect on u is defined by

$$\Delta_h u : x \longmapsto u(x + h) - u(x) \qquad x \in R.$$

With this notation, Equation (1.6) in *S: page 7* can be rewritten concisely as

$$u'(x) \simeq \frac{1}{h}\{\Delta_h u(x)\}.$$

This is hardly surprising since we originally defined $u'(x)$ (*Unit M100 12, Differentiation I*) as

$$\lim_{h \to 0} \frac{1}{h}\{\Delta_h u(x)\}.$$

However, it is pointed out in *S: pages 6 and 7* that the central-difference approximation is more accurate. We therefore introduce the **central-difference operator** δ_h on the set of functions $R \longrightarrow R$, whose effect on u is given by

$$\delta_h u : x \longmapsto u(x + \tfrac{1}{2}h) - u(x - \tfrac{1}{2}h).$$

We may compose the central-difference operator with itself to obtain $\delta_h^2 = \delta_h \circ \delta_h$, etc. Clearly

$$\delta_h^2 u : x \longmapsto u(x + h) - 2u(x) + u(x - h),$$

so that we may rewrite Equation (1.4) in *S: pdge 6* as

$$u''(x) \simeq \frac{1}{h^2}\{\delta_h^2 u(x)\}.$$

If we assume that the function u is tabulated at intervals of h, we encounter some difficulty in expressing Equation (1.5) in *S: page 6* in terms of the central-difference operator δ_h since $\delta_h u(x)$ involves values of u at the untabulated points $x \pm \frac{1}{2}h$. We resolve this problem by introducing the **averaging operator** μ_h, whose effect on u is given by

$$\mu_h u : x \longmapsto \tfrac{1}{2}[u(x + \tfrac{1}{2}h) + u(x - \tfrac{1}{2}h)].$$

We now have

$$\mu_h \circ \delta_h u : x \longmapsto \tfrac{1}{2}[\delta_h u(x + \tfrac{1}{2}h) + \delta_h u(x - \tfrac{1}{2}h)] = \tfrac{1}{2}[u(x + h) - u(x - h)],$$

and so Equation (1.5) becomes, with our notation,

$$u'(x) \simeq \frac{1}{h}\{\mu\delta_h u(x)\}.$$

Note the convention that the subscript on μ and the mapping product symbol are omitted when the meaning is clear from the context.

When dealing with two independent variables x and t with, perhaps, different constant spacings h and k in the two directions, it is more convenient to suppress the subscripts h and k and write, for example,

$$\Delta_x u(x_i, t_j) = u(x_i + h, t_j) - u(x_i, t_j)$$

$$\delta_x^2 u(x_i, t_j) = u(x_i + h, t_j) - 2u(x_i, t_j) + u(x_i - h, t_j).$$

The subscript x here denotes *differencing in the x-direction.*

With this notation as a useful shorthand, we can write Equations (1.8), (1.9) and (1.10) in ***S**: page 8* as

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \simeq \frac{1}{h^2}\,\delta_x^2 u_{i,j},$$

$$\left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} \simeq \frac{1}{k^2}\,\delta_t^2 u_{i,j},$$

and

$$\left(\frac{\partial u}{\partial t}\right)_{i,j} \simeq \frac{1}{k}\,\Delta_t u_{i,j},$$

respectively.

These approximations have local errors, and we might wish occasionally to use more accurate formulas. The derivation of such formulas is quite lengthy and we shall not deal with it here. However, we may, from time to time, quote results which you will have to accept on trust. (There is a good account of the theory involved in F. B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill, 1956.)

5.2 EXPLICIT METHODS OF SOLUTION FOR PARABOLIC AND HYPERBOLIC EQUATIONS

READ S: the section entitled **An explicit method of solution,** *pages 10 to 17, omitting page 14, lines −18 to −7.*

We have not asked you to read the passage in *S: pages 9 and 10* which shows that the one-dimensional heat equation for a finite rod can be expressed in terms of non-dimensional variables as

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad x \in (0, 1), t > 0.$$

Notes

(i) *S: page 10, Equation (2.4)*

We consider r to be a parameter whose value we can choose to make the formula relating the various values of u at the mesh points (or, as we shall call it, the **finite-difference scheme**) as useful as possible. Another piece of jargon that we sometimes use is that we call the equation in *S: page 10, line −5* the **finite-difference replacement** of the partial differential equation (2.3). The parameter r is known as the **mesh ratio** of the scheme.

(ii) *S: page 11, line 3*
It is usual to refer to a **time level** rather than a time row. The process of calculating values along one time level from the values on the previous time level(s) is called *stepping forward in time* where the step is of length k (the mesh length in the t-direction).

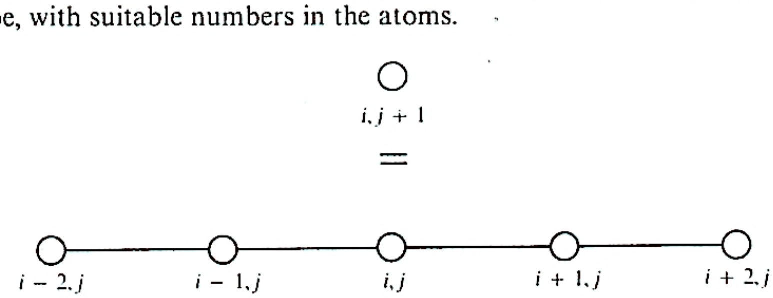
(iii) *S: page 11, Example 2.1*
Equation (2.4) on *S: page 10* is valid only for points in the interior of the solution domain. We cannot make use of this formula to evaluate the solution at boundary points. However, we are given the boundary values along $x = 0$ and $x = 1$ and we are able to evaluate the solution at all points along a time level.

(iv) *S: page 12, line 8*
The choices for h and k (and hence r) are arbitrary at this stage. The only information we have about choosing values for h and k comes from our discussion of Taylor's Theorem, from which we may conclude that "small" absolute values of h and k would be needed to give good results.

Note also that there is an infinitude of choices for h and k which would yield the same value for r. For example, taking $h = 1$ and $k = \frac{1}{10}$ also yields $r = \frac{1}{10}$.

(v) *S: page 12, Fig. 2.2*
The idea of a **molecule** representing a finite-difference scheme is a useful one. Each **atom** represents the value of u at a single mesh point. The whole molecule gives a summary of the coefficients of a scheme and its form. (Note that a *single* atom at a **future** time level, i.e. above the "equals" sign, indicates an *explicit* scheme.)

It also tells us how many time levels are involved, and how many initial and/or boundary conditions are required in order to use the scheme. This last point can be illustrated by considering an explicit scheme (on some level) of the following type, with suitable numbers in the atoms.



(Note that we have returned the "equals" sign to its familiar orientation. The interpretation of the molecule remains the same.)

This, for example, could be produced by using, in place of the simple formula

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \simeq \frac{1}{h^2}\,\delta_x^2 u_{i,j},$$

the higher-order formula

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \simeq \frac{1}{h^2}\,\{\delta_x^2 u_{i,j} - \tfrac{1}{12}\delta_x^4 u_{i,j}\}.$$

(You need not check this formula.)

Because of the presence of *two* atoms to the right and left of i, j this scheme requires two conditions on each of the boundaries. Of course, we *have* these two conditions; namely, we are given

(a) $\quad u(0, t) = 0 \qquad t \geqslant 0,$

say, and since this implies that

$$\frac{\partial u}{\partial t}(0, t) = 0 \qquad t \geqslant 0,$$

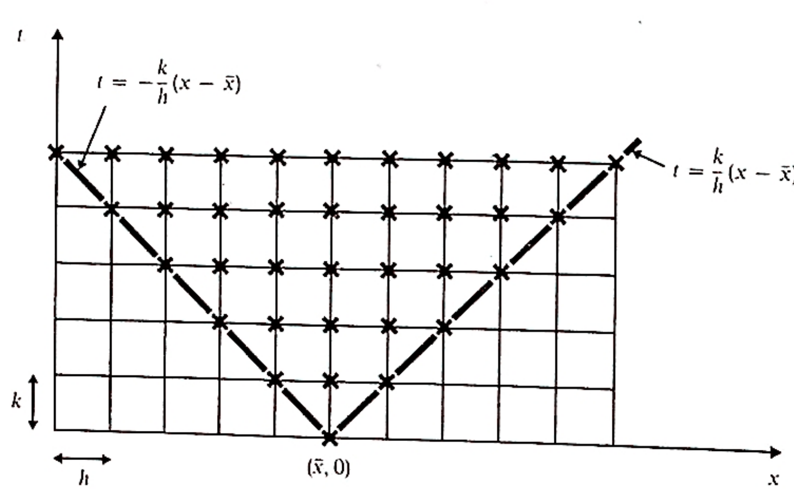
we have, from the differential equation,

(b) $\quad \dfrac{\partial^2 u}{\partial x^2}(0, t) = 0 \qquad t \geqslant 0.$

Naturally the choice of finite-difference replacement cannot alter the number of boundary conditions required for the partial differential equation to be properly posed.

(vi) *S: page 13, line −6*
We shall ask you to obtain this solution in SAQ 21 of *Unit 6, Fourier Series.*

(vii) *S: page 14, lines 1 to 8*
The *finite-difference* scheme allows the effect of the discontinuity to propagate as an error in the solution. We can see this as follows. The points marked with a cross in the diagram are those points at which the numerical solution to the differential equation, using the finite-difference scheme (2.6) in *S: page 12*, depends on the value of u at $(\bar{x}, 0)$. We call this set of points the **numerical domain of influence** of the point $(\bar{x}, 0)$.



Now, for the *differential* equation which we are considering there is but one family of characteristics, t = constant, as we have seen in *Unit 2, Classification and Characteristics*. As a result, the discontinuity in $\partial u/\partial x$ at $(\frac{1}{2}, 0)$ does not cross any characteristic, and so does not propagate.

On the other hand, we see that, for given h and k, the domain of influence of $(\bar{x}, 0)$ in the finite-difference scheme is bounded below by the lines

$$t = \frac{k}{h}(x - \bar{x}) \quad \text{and} \quad t = -\frac{k}{h}(x - \bar{x}),$$

along which the error propagates. (The speed of propagation is seen to be h/k.) For fixed h, as k approaches zero (and hence also $r \rightsquigarrow 0$) these two numerical characteristics both approach

$$t = 0,$$

i.e., the characteristic of the differential equation which passes through $(\bar{x}, 0)$.

In fact, even if h is not kept fixed, but h and k both approach zero in such a way that $r = k/h^2$ remains constant, $k/h = hr$ still approaches zero and the characteristics of the numerical scheme approach $t = 0$.

The error in our computation is generated by using a Taylor approximation to $\partial^2 u/\partial x^2$ at $(\frac{1}{2}, 0)$—a point where this second derivative does not exist! However, for our particular choice of h and k the scheme is stable with respect to this error. As the results in Table 2.4 show, the magnitude of the error decreases as t increases.

The second paragraph of *S: page 14* is referring to the *numerical* solutions of parabolic equations. As we shall not be investigating this effect any further there is no necessity to refer to Exercise 12 in *S: Chapter 3* as suggested in line 8.

(viii) *S: page 15, Table 2.6*
The percentage error column of Table 2.6 indicates larger errors than in Table 2.3 (*S: page 13*). This is the result of two effects.

The first is that in Case 2 the rate of propagation of the error generated by the discontinuity is one fifth of that in Case 1, since now $h/k = 20$ rather than 100. We should therefore expect the results to be poorer until this effect has subsided. The second effect is that since k is much larger than before the local truncation error is larger.

(ix) *S: page 16, lines 1 to 4*
The values occurring in Table 2.7 which are less than 0 or greater than 1 obviously contradict the maximum principle for the heat equation (*W: page 60, lines −2 and −1*). This gives us a warning that something is going wrong.

(x) *S: page 16, line −3*
Smith describes a finite-difference replacement of a differential equation as **valid** if it is convergent, stable and compatible. These are topics covered in *S: Chapter 3* and we shall discuss them in *Unit 8*. A short discussion of stability in relation to the explicit scheme of this section is presented in the following text.

General Comment

A possible reason for the failure of the numerical scheme in Case 3 arises from the fact that we are performing a step-by-step process with a recurrence relation (with two independent arguments i and j). We already know, from the work on recurrence relations with one argument (*Unit M201 7*), that such a computation can sometimes give rise to serious induced instability.

In Table 2.7 (*S: page 15*) the instability does not arise through local *rounding error* because the entries in this table were computed with exact arithmetic as you can easily verify by hand. There is, however, another significant component of local error, present in all these computations. It is the truncation error which comes from the fact that Equation (2.4) (*S: page 10*) is not an *exact* replacement of the differential equation. Remember that each of the finite-difference formulas obtained in Section 5.1 was subject to a local truncation error as we saw in note (i) of that section. The finite-difference equation (2.4), therefore, contains a local error which is a sum of terms $O(h^2)$ and $O(k)$ appearing in Equations (1.8) and (1.10) of *S: page 8*.

Formally, we define the **local truncation error at a point** of the finite-difference replacement of a differential equation as the error it introduces in evaluating the solution at that point, i.e. the terms truncated in replacing derivatives by their Taylor approximations. For example, let U denote the true solution of

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}$$

subject to the stated subsidiary conditions. The local truncation error at $i, j+1$ is

$$T_{i,j} = \frac{1}{k}(U_{i,j+1} - U_{i,j}) - \frac{1}{h^2}(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}).$$

If we can find the local truncation error at each point it will give us an indication of how well the true solution of the partial differential equation satisfies the finite-difference equation. In the present case we already know the true solution of the differential equation. Therefore, we have substituted the values of the true solution into the finite-difference equation for each of the three cases and have recorded the value of the local truncation error, $T_{i,j}$, at selected mesh points. The results are shown in tables for the cases $r = 0.1$, 0.5 and 1.0 respectively.

We can see from these results the different values of $T_{i,j}$ at any particular point in the three cases. For example, look at the figures corresponding to $t = 0.02$ ($j = 20$ in Case 1, $j = 4$ in Case 2 and $j = 2$ in Case 3). We can account for these differences by the different sizes of k, in view of the $O(k)$ terms in the local truncation error.

These results do not, by themselves, account for the drastic behaviour in Case 3. We can conclude that the effects of truncation errors made at each step are accumulating in the third case as the step-by-step process progresses. In the first two cases these errors must be diminishing in their effect. The only difference between the cases is the value of r and we conclude that it must be this which governs the stability or instability of the scheme.

The large errors for $j = 0$, especially in $T_{5,0}$, are due to the discontinuity in $\partial u/\partial x$ at $(\frac{1}{2}, 0)$. The fact that, in the first two cases, this error does not propagate through the solution illustrates dramatically how the effect of the local errors diminishes as t increases, for these cases.

It is also interesting to point out that had Case 3 been computed on a digital computer where rounding errors are introduced, these too would have accumulated. However, rounding errors are often smaller than truncation errors, and in this unit we concentrate on the latter.

Note that in this section we have discussed only the parabolic equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

There is an explicit scheme for the hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

which may be developed along the lines discussed here. It has a similar property that its stability depends on its mesh ratio. This method is the subject of SAQ 1.

Case 1: $h = 0.1$, $k = 0.001$, $r = 0.1$

		$i = 1$	2	3	4	5
		$x = 0.1$	0.2	0.3	0.4	0.5
($j = 0$)	$t = 0.000$	2.23517E − 05	2.08616E − 04	1.04308E − 04	−7.88033E − 01	−3.13647E + 01
1	0.001	−4.47035E − 04	2.60770E − 04	2.50787E − 02	1.62046	−3.67489
2	0.002	2.23517E − 05	1.70246E − 03	1.77227E − 01	6.36525E − 01	−1.63817
3	0.003	1.49012E − 04	1.40928E − 02	2.45437E − 01	2.24337E − 01	−9.68307E − 01
4	0.004	1.08778E − 03	3.46191E − 02	2.37897E − 01	5.45979E − 02	−6.55293E − 01
5	0.005	3.72529E − 03	5.29960E − 02	2.03758E − 01	−2.01315E − 02	−4.80771E − 01
	⋮					
10	0.01	2.46763E − 02	6.81467E − 02	6.34789E − 02	−7.27028E − 02	−1.78874E − 01
	⋮					
20	0.02	2.04742E − 02	2.42516E − 02	−1.14739E − 03	−4.32655E − 02	−6.47008E − 02

Case 2: $h = 0.1$, $k = 0.005$, $r = 0.5$

		$i = 1$	2	3	4	5
		$x = 0.1$	0.2	0.3	0.4	0.5
($j = 0$)	$t = 0.000$	−5.60284E − 04	−3.05414E − 02	−6.79231E − 01	−6.66516	8.08468
1	0.005	−6.26862E − 02	−3.49882E − 01	−6.72478E − 01	3.26108E − 01	1.53010
2	0.010	−1.39344E − 01	−2.92724E − 01	−1.93632E − 01	3.33082E − 01	6.92152E − 01
3	0.015	−1.26433E − 01	−1.75536E − 01	−3.60668E − 02	2.54286E − 01	4.13870E − 01
4	0.020	−8.60870E − 02	−9.31948E − 02	2.13026E − 02	1.96314E − 01	2.82646E − 01
5	0.025	−4.97580E − 02	−4.05758E − 02	4.50074E − 02	1.57166E − 01	2.09284E − 01
	⋮					
10	0.050	1.50621E − 02	3.53306E − 02	5.99980E − 02	8.13424E − 02	8.98718E − 02

Case 3: $h = 0.1$, $k = 0.01$, $r = 1$

		$i = 1$	2	3	4	5
		$x = 0.1$	0.2	0.3	0.4	0.5
($j = 0$)	$t = 0.00$	−3.89755E − 02	−3.44896E − 01	−2.01015	−7.98562	1.74325E + 01
1	0.01	−3.42774E − 01	6.35606E − 01	−3.44312E − 01	7.68459E − 01	1.48830
2	0.02	1.85341E − 01	1.88848E − 01	7.38025E − 02	4.58241E − 01	6.44755E − 01
3	0.03	−4.47214E − 02	2.23517E − 03	1.42622E − 01	3.10755E − 01	3.85380E − 01
4	0.04	1.68860E − 02	6.75470E − 02	1.53589E − 01	2.38597E − 01	2.74277E − 01
5	0.05	3.94821E − 02	8.97855E − 02	1.48582E − 01	1.98489E − 01	2.18272E − 01

SAQ 1

Write down an explicit finite-difference scheme, using central differences, for the hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

Draw the molecule for your scheme. If the mesh length is h in the x-direction and k in the t-direction what is the mesh ratio (compare with note (i) of this section)? Write down the numerical domain of influence of some general point $(\bar{x}, \bar{t})$, $\bar{t} \geqslant 0$, for your scheme. What initial data does your scheme require?

(Solution on p. 32.)

In SAQ 2, which follows, parts (a) to (f) are designed to assist you in constructing a computer program which will yield the numerical solution to a particular case of the diffusion equation. If you are short of time then make use of the library program $EXP321, which has been developed as a solution to this problem, and tackle part (g) only.

SAQ 2

(a) Write a BASIC statement which can be used to evaluate $u_{i,j+1}$ according to Equation (2.4) of *S: page 10*. Assume that $u_{i,j}$, $u_{i-1,j}$, $u_{i+1,j}$ and r are already available in the computer.

(b) Place the solution to (a) within a BASIC loop which will calculate $u_{i,j+1}$ for $i = 2, 3, \ldots, N-1$ for some j, where N has already been read into the computer. Assume that $u_{i,j} (i = 1, 2, \ldots, N)$ are already available in the computer.

(c) Incorporate the result of (b) into another BASIC loop which will calculate $u_{i,j+1}$ for $j = 1, 2, \ldots, M$, where M is already available in the computer. Assume that $u_{1,j}$ and $u_{N,j} (j = 1, 2, \ldots, M)$, and $u_{i,1} (i = 2, 3, \ldots, N-1)$ are already available in the computer.

(d) In (c) what do the values

(i) $u_{1,j}$ and $u_{N,j}$ $\quad j = 1, 2, \ldots, M,$
(ii) $u_{i,1}$ $\quad i = 1, 2, \ldots, N$

represent?

(e) Why are zero values not used for the subscripts i and j in the previous parts of this question?

(f) Write a complete BASIC program which will solve the initial-boundary value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad 0 < x < \pi, \quad t > 0$$

subject to the initial condition

$$u(x, 0) = \sin x \qquad 0 \leqslant x \leqslant \pi$$

and the boundary conditions

$$u(0, t) = u(\pi, t) = 0 \qquad t \geqslant 0.$$

Let N, the number of mesh points in the x-direction, and k, the mesh length in the t-direction, be read in as data, and arrange for r to be printed out. You may assume that the solution is not required for $t > 20k$. The output from your program should show, at each mesh point:

the analytical solution to the problem, which is given by

$$u(x, t) = e^{-t} \sin x \qquad (x, t) \in [0, \pi] \times R_0^+;$$

the solution to the finite-difference scheme;

the error, i.e. the difference between the two solutions.

To avoid an inordinate amount of output show the analytical solution, numerical solution and error at only one value of x for each time step. Choose that value of x where you expect the error to be greatest.

(g) Run the program developed in part (f) with various values of N and k to investigate the nature of the numerical solution for various values of the mesh ratio, r, both greater and smaller than $\frac{1}{2}$. Check your results by comparison with those obtained from the library program $EXP321.

(Solution on p. 32.)

5.3 AN IMPLICIT METHOD OF SOLUTION

5.3.1 The Crank–Nicolson Method

READ S: the sections entitled **Crank-Nicolson implicit method,** *pages 17 to 20, and* **A weighted average approximation,** *pages 23 and 24, omitting the final reference to Chapter 3.*

Notes

(i) *S: page 17, line 18*
You may recall that we used this averaging device in Section 21.2.2 of *Unit M201 21* to obtain the trapezoidal-rule formula for ordinary differential equations, and in Section 2.3 of *Unit 2, Classification and Characteristics.* The new formula has the property that the $O(k)$ term of the truncation error for the explicit formula becomes $O(k^2)$ as we shall see in Section 5.5.

(ii) *S: page 18, lines −5 and −4*
Since $k = rh^2$ we must be careful not to choose too large a value for r, since the resulting large value for k would produce a large local truncation error.

(iii) *S: page 18, line −3*
The advantage in choosing $r = 1$ is that it simplifies the numerical formula and hence reduces the number of calculations required. This situation is of use in both hand and computer calculation since it reduces the total time required to obtain the solution. Note that we can still control the accuracy of the scheme with $r = 1$ by choosing an appropriate value for h (in which case $k = h^2$).

(iv) *S: page 19, lines 5 to 9*
Note the matrix form of these equations:

$$\begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & 4 & -1 & \\ & & -1 & 4 & -1 \\ & & & -2 & 4 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.8 \\ 1.2 \\ 1.6 \\ 1.6 \end{bmatrix},$$

which is of the form $A\mathbf{u} = \mathbf{b}$ where A is **tridiagonal**, that is, its nonzero elements appear on and adjacent to its main diagonal.*

(v) *S: page 23, lines −6 to −1*
This is an example of a technique which is used widely in numerical mathematics. Having obtained a useful formula (in this case the Crank–Nicolson method) we try to generalize it to a formula which reduces to other well-known formulas for particular values of some parameter (in this case θ). Any analysis which needs to be applied to each of the particular formulas now needs only to be applied to the one general formula and the results for the particular formulas are obtained by substituting in the relevant value of the parameter.

General Comment

Once again we have restricted ourselves to a parabolic equation. Implicit methods can be found for hyperbolic equations. For example, for the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2},$$

* Such a matrix is called *triple diagonal* in *Unit M201 30, Numerical Solution of Eigenvalue Problems.*

an analogue to the weighted average approximation in the parabolic case is

$$\frac{\delta_t^2 u_{i,j}}{k^2} = \frac{\theta\delta_x^2 u_{i,j+1} + (1-2\theta)\delta_x^2 u_{i,j} + \theta\delta_x^2 u_{i,j-1}}{h^2}.$$

SAQ 3

(a) Draw the molecule for the Crank–Nicolson method with mesh ratio r.

(b) Write the Crank–Nicolson implicit method in terms of the central-difference operator δ_x^2.

(Solution on p. 35.)

SAQ 4

Write a section of a BASIC program which will calculate the matrix of coefficients of the variables $u_{i,j+1} (i = 2, 3, \ldots, N-1)$ generated by the Crank–Nicolson implicit method with general r. Let $u_{1,j}$ and $u_{N,j} (j = 1, 2, \ldots, M)$ be known boundary values and store the elements of the matrix in the array C.

(Solution on p. 35.)

SAQ 5

Write down the matrix of coefficients for the Crank–Nicolson method applied to a problem involving six internal mesh points along each time level. Assume that the boundary values $u_{0,j+1}$ and $u_{7,j+1}$ are known. What special form does this matrix have?

(Solution on p. 36.)

The library program $CNH321 uses the Crank–Nicolson method to solve the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad a < x < b, \quad t > 0,$$

with the initial condition

$$u(x, 0) = \sin \pi x \qquad a \leqslant x \leqslant b$$

and the boundary conditions

$$u(a, t) = 0 \qquad t > 0,$$

$$u(b, t) = 0 \qquad t > 0.$$

The program is written in BASIC. You can replace statements 160, 170 or 180 by new DEF statements to specify alternative initial and boundary conditions for this problem. For example, the initial condition

$$u(x, 0) = x^2 + 6x + 1 \qquad a \leqslant x \leqslant b$$

requires that statement 160 be

```
160 DEF FNI(X) = X * X + 6 * X + 1.
```

Input data consists of:

(i) the values a, b defining the boundaries ($x = a$ and $x = b$) of the domain;
(ii) the number of mesh points along each time-level;
(iii) the value of the mesh ratio;
(iv) the total number of steps to be taken in the t-direction.

A listing of $CNH321 follows. The section between statements 400 and 500 solves a tridiagonal system of equations by the method to be explained in Section 5.3.2.

```
10  PRINT "THIS PROGRAM USES THE CRANK-NICOLSON METHOD TO GIVE A "
2Ø  PRINT "NUMERICAL SOLUTION OF THE HEAT CONDUCTION EQUATION."
3Ø  PRINT "THE INITIAL CONDITIONS ARE SPECIFIED BY A DEF STATEMENT "
4Ø  PRINT "AT LINE 16Ø."
5Ø  PRINT "BOUNDARY CONDITIONS ARE SPECIFIED BY DEF STATEMENTS IN "
6Ø  PRINT "LINE 17Ø(FNA-B.V. AT X=A)AND LINE 18Ø (FNB-B.V. AT X=B)."
7Ø  PRINT
8Ø  DIM A[2Ø],B[2Ø],C[2Ø],D[2Ø],G[2Ø],U[2Ø],W[2Ø]
9Ø  REM***INPUT SEQUENCE***
1ØØ  PRINT "INPUT END POINTS OF DOMAIN-X=A AND X=B";
11Ø  INPUT A,B
12Ø  PRINT "INPUT NUMBER OF MESH POINTS,N";
13Ø  INPUT N
14Ø  H=(B-A)/(N-1)
15Ø  PRINT "INPUT PARAMETER R";
155  INPUT R
16Ø  DEF FNI(X)=SIN(3.14159*X)
17Ø  DEF FNA(T)=Ø
18Ø  DEF FNB(T)=Ø
19Ø  PRINT "INPUT THE NUMBER OF TIME STEPS REQUIRED";
2ØØ  INPUT M
21Ø  PRINT
22Ø  REM***SET UP COEFFICIENTS OF LINEAR EQUATIONS***
23Ø  FOR I=1 TO N
24Ø  A[I]=R
25Ø  B[I]=2+2*R
26Ø  C[I]=R
27Ø  NEXT I
28Ø  GOSUB 59Ø
29Ø  PRINT "                              SOLUTION"
3ØØ  PRINT "                             -----------"
31Ø  FOR J=Ø TO M
32Ø  GOSUB 65Ø
33Ø  PRINT "TIME=";T
34Ø  PRINT "-----------"
35Ø  IF J=Ø THEN 51Ø
36Ø  REM***CALCULATE R.H.S. OF LINEAR EQUATIONS***
37Ø  FOR I=2 TO N-1
38Ø  D[I]=R*U[I-1]+(2-2*R)*U[I]+R*U[I+1]
39Ø  NEXT I
4ØØ  REM*SOLUTION OF TRIDIAGONAL SYSTEM*
41Ø  W[1]=Ø
42Ø  G[1]=U[1]
43Ø  FOR I=2 TO N-1
44Ø  W[I]=C[I]/(B[I]-A[I]*W[I-1])
45Ø  G[I]=(D[I]+A[I]*G[I-1])/(B[I]-A[I]*W[I-1])
46Ø  NEXT I
47Ø  REM*CALCULATE THE U(I)*
48Ø  FOR I=N-1 TO 2 STEP -1
49Ø  U[I]=W[I]*U[I+1]+G[I]
5ØØ  NEXT I
51Ø  REM***OUTPUT OF RESULTS FOR ONE TIME STEP***
52Ø  FOR I=1 TO N
53Ø  PRINT U[I],
54Ø  NEXT I
55Ø  PRINT
56Ø  PRINT "                        ---------------"
57Ø  NEXT J
58Ø  STOP
59Ø  REM***SUBROUTINE FOR INITIAL CONDITIONS ALONG T=Ø***
```

```
600  FOR I=2 TO N-1
610  X=A+(I-1)*H
620  U[I]=FNI(X)
630  NEXT I
640  RETURN
650  REM***SUBROUTINE FOR B.C.'S ALONG X=A AND X=B***
660  T=J*H*H*R
670  U[1]=FNA(T)
680  U[N]=FNB(T)
690  RETURN
700  END
```

SAQ 6

Use the library program $CNH321 to obtain the solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad 0 < x < 1, \quad t > 0$$

subject to

$$u(x, 0) = \sin \pi x \qquad 0 \leqslant x \leqslant 1$$

and

$$u(0, t) = u(1, t) = 0 \qquad t \geqslant 0,$$

at the point (0.5, 0.1). You should use a mesh length $h = 0.1$ and a mesh ratio $r = 0.5$.

(Solution on p. 36.)

5.3.2 The Solution of Tridiagonal Systems of Equations

Although Gauss elimination with pivoting is a stable and natural method for solving sets of linear equations (*Unit M201 8*), the simple tridiagonal form of the system of equations generated by the Crank–Nicolson implicit method (see SAQ 5) allows a modification of the method which is simple to apply in practice.

READ S: the section entitled **Solution of the equations by Gauss's elimination method,** *pages 20 to 23.*

General Comment

The method of solution in the text is equivalent to Gauss elimination, but it turns out that no interchanges are necessary to ensure that the multipliers in the elimination are $\leqslant 1$ in absolute value. As we showed in *Unit M201 8*, the elimination method is then stable. We should, however, prove the assertion that in our case the multipliers are all $\leqslant 1$ automatically, so that the pivots always lie on the diagonal of the coefficient matrix.

In order to show that the method of solution of the equations in *S: page 20* is stable we shall show two things.

(a) The multipliers in the process have absolute value $\leqslant 1$, provided

$$a_i > 0, b_i > 0, c_i > 0 \text{ and } b_i > a_{i+1} + c_{i-1}$$

for each value of i. (We define $c_0 = a_N = 0$.)

(b) If, in addition, $b_i > a_i + c_i$ for each value of i, then the back-substitution is stable. (We define $a_1 = c_{N-1} = 0$.)

The final condition in (a) is that each diagonal element exceeds the sum of the absolute values of the other elements in the same column. The condition in (b) is similar with "row" instead of "column". For symmetric matrices these two conditions are clearly equivalent.

The matrix of coefficients for the Crank–Nicolson method satisfies all the conditions, since for each i

$$a_i = r,$$

$$b_i = 2 + 2r,$$

and

$$c_i = r.$$

If you are short of time, you should omit the following proof, since it digresses from the main theme of the unit.

Proof

(a) From the first pair of equations in ***S****: page 21* we see that the multiplier m_i used at this stage to eliminate u_{i-1} is just a_i/α_{i-1}, and from the formula for α_i below Equation (2.12) we easily find that the multipliers $m_i = a_i/\alpha_{i-1}$ satisfy the recurrence relation

$$m_{i+1} = \frac{a_{i+1}}{b_i - c_{i-1}m_i} \qquad i = 2, 3, \ldots, N-1,$$

with

$$m_2 = \frac{a_2}{b_1}.$$

We want to show that $|m_{i+1}| < 1$ for $i = 1, 2, \ldots, N-2$, provided that

$$a_i > 0, b_i > 0, c_i > 0, \text{ and } b_i > a_{i+1} + c_{i-1}.$$

We have

$$0 < m_2 = \frac{a_2}{b_1} < 1 \quad \text{since} \quad b_1 > a_2 > 0.$$

Next

$$0 < m_3 = \frac{a_3}{b_2 - c_1 m_2} \quad \text{since } a_3 > 0, b_2 > c_1 \text{ and } m_2 < 1$$

$$< \frac{a_3}{b_2 - c_1} \quad \text{since } m_2 < 1 \text{ and } c_1 > 0$$

$$< \frac{a_3}{(a_3 + c_1) - c_1} = 1 \quad \text{since } b_2 > a_3 + c_1$$

and, by induction, it may be shown that

$$0 < m_i < 1.$$

(b) In *Unit M201 8* it was asserted, but not proved, that if all $|m_i| \leqslant 1$ then the back-substitution is also stable. We can in fact prove this easily for this special case, when the matrix is tridiagonal and no interchanges are needed in the elimination.

The required solution is computed from the *backwards* recurrence relation (***S****: page 21, line −1*)

$$u_i = \frac{1}{\alpha_i}(S_i + c_i u_{i+1}),$$

where the S_i and α_i are given by the equations at the top of ***S****: page 22*. Now in Section 7.3.2 of *Unit M201 7* (on computations with recurrence relations) we saw that the

successive determination of $u_{n-1}, u_{n-2}, \ldots, u_1$ from a recurrence relation of the form

$$u_{i-1} = p_i u_i + q_i,$$

with u_n specified, suffers neither from inherent nor from induced instability if $|p_i| < 1$ for $i = n, n-1, \ldots, 2$.

Here $p_i = c_{i-1}/\alpha_{i-1}$, and from the defining equations for α_i we deduce the recurrence relation

$$p_{i+1} = \frac{c_i}{b_i - a_i p_i} \qquad i = 2, 3, \ldots .$$

with

$$p_2 = \frac{c_1}{b_1}.$$

Thus, using the given conditions,

$$0 < p_2 = \frac{c_1}{b_1} < 1 \qquad \text{since } b_1 > c_1 > 0.$$

Next,

$$0 < p_3 = \frac{c_2}{b_2 - a_2 p_2} \qquad \text{since } c_2 > 0,\ b_2 > a_2 \text{ and } p_2 < 1$$

$$< \frac{c_2}{b_2 - a_2} \qquad \text{since } p_2 < 1 \text{ and } a_2 > 0$$

$$< \frac{c_2}{(a_2 + c_2) - a_2} = 1 \qquad \text{since } b_2 > a_2 + c_2,$$

and the general result, $0 < p_i < 1$, follows by induction.

The method is therefore completely stable, and though the arithmetic is identical with that of Gaussian elimination, the recurrence relations in ***S**: page 21, line −2 to page 22, line 3* express the method in a convenient and compact form for programming purposes.

SAQ 7

Calculate the solution to Example 2.2 of ***S**: page 18* by the fully implicit backwards time-difference method (i.e. the weighted average approximation with $\theta = 1$) utilizing the recurrence relation version of the Gauss elimination method for the resulting tridiagonal system of equations. Take $h = \frac{1}{10}$, and $r = 1$ and perform the calculations for the first time step only.

(Solution on p. 36.)

SAQ 8

Write a BASIC program to solve a tridiagonal system of linear equations by the recurrence method. The input should include

(i) the number of equations to be solved,
(ii) the nonzero coefficients, and
(iii) the elements on the right-hand sides of the equations.

(Solution on p. 37.)

SAQ 9

Solve the equations

$$\begin{aligned} 2x_1 - \tfrac{3}{2}x_2 &= -1 \\ -\tfrac{3}{2}x_1 + 2x_2 - \tfrac{3}{2}x_3 &= -2 \\ -\tfrac{3}{2}x_2 + 2x_3 - \tfrac{3}{2}x_4 &= -3 \\ -\tfrac{3}{2}x_3 + 2x_4 &= \tfrac{7}{2} \end{aligned}$$

by the recurrence method for tridiagonal systems using

(i) exact arithmetic (use fractions),
(ii) the library program STRI321 (see solution to SAQ 8).

(Solution on p. 38.)

SAQ 10

Are the equations produced by the implicit scheme for hyperbolic equations (see General Comment in Section 5.3.1), for $\theta > 0$, soluble without instability by the method of Gauss elimination without pivoting?

(Solution on p. 39.)

5.4 DERIVATIVE INITIAL AND BOUNDARY CONDITIONS

5.4.1 Initial Conditions

Consider the following initial value problem:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \qquad 0 < x < 1, \quad t > 0$$

$$u(x, 0) = f(x) \qquad 0 \leqslant x \leqslant 1$$

$$\frac{\partial u}{\partial t}(x, 0) = g(x) \qquad 0 \leqslant x \leqslant 1.$$

Suppose we use the explicit finite-difference scheme

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

to solve the initial value problem. We can write the numerical scheme as

$$u_{i,j+1} = 2(1 - p^2)u_{i,j} + p^2(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1}$$

where the mesh ratio $p^2 = k^2/h^2$. We see immediately that to calculate values of u along $t = (j + 1)k$ we need to know the values of u along $t = jk$ and $t = (j - 1)k$. The implication of this is that we shall need to specify initial values of u along both $j = 0$ and $j = 1$.

We already have $u(x, 0) = f(x)$ given along $t = 0$ so we use the other condition, namely

$$\frac{\partial u}{\partial t}(x, 0) = g(x),$$

to give us values along $t = k$. This derivative initial condition can be treated by the method of Section 5.1, that is, we replace it by a finite-difference formula. The choice of the forward-difference formula,

$$\frac{\partial u}{\partial t}(x_i, 0) = \left(\frac{\partial u}{\partial t}\right)_{i,0} \simeq \frac{u_{i,1} - u_{i,0}}{k},$$

yields the expression

$$u_{i,1} - u_{i,0} = kg_i$$

where g_i denotes $g(x_i)$. Now $u_{i,0} = u(x_i, 0) = f(x_i)$, by the first initial condition, and this allows us to write

$$u_{i,1} = f_i + kg_i \qquad i = 0, 1, \ldots$$

where f_i denotes $f(x_i)$. Thus we have obtained values along the time level $t = k$ as well as along the initial line $t = 0$ and we can use the explicit finite-difference scheme to obtain the solution for $t > k$.

5.4.2 Boundary Conditions

Read S: the section entitled **Derivative boundary conditions,** *pages 32 to 40, ignoring the references to Chapter 3.*

Notes

(i) *S: page 33, line −6*
Beware! Here h is the positive physical constant introduced in *S: page 33, line 6*, and *not* the mesh spacing in the x-direction, which is now designated by δx.

(ii) *S: page 35, lines −10 and −9*
u is symmetric about $x = \frac{1}{2}$ if and only if

$$u(x, t) = u(1 - x, t) \qquad 0 \leqslant x \leqslant 1, \quad t \geqslant 0.$$

You may verify that if the function u satisfies the problem of Example 2.4 then so does the function

$$(x, t) \longmapsto u(1 - x, t) \qquad 0 \leqslant x \leqslant 1, \quad t \geqslant 0.$$

Since the solution to the problem is unique, it must be symmetric about $x = \frac{1}{2}$.

(iii) *S: page 35, line −2*
Since the solution is symmetric we obtain $u_{0,j} = u_{10,j}$, $u_{1,j} = u_{9,j}$, $u_{2,j} = u_{8,j}$, $u_{3,j} = u_{7,j}$ and $u_{4,j} = u_{6,j}$, and we need only solve for the six unknowns u_i ($i = 0, 1, \ldots, 5$). The six equations required are those for $u_{i,j+1}$ ($i = 0, 1, 2, 3, 4$) plus that for $u_{5,j+1}$ with $u_{6,j}$ replaced by $u_{4,j}$.

SAQ 11

Why is $\partial u/\partial x$ represented more accurately by a central-difference formula than a forward-difference formula, as claimed in *S: page 34, lines 3 and 4*?

(Solution on p. 40.)

SAQ 12

S: page 50, Exercise 5

(Solution on p. 40.)

You should omit SAQ 13 if you are short of time.

SAQ 13

S: page 51, Exercise 6
If you are calculating by hand, compute u on just the first time level.

(Solution on p. 40.)

5.5 ORDER OF LOCAL TRUNCATION ERROR

In Section 5.2 we computed the local truncation error of a simple explicit method applied to the heat equation in one (space) dimension. That is, we found the extent to which the true solution of the differential equation did not satisfy the finite difference equation by computing

$$T_{i,j} = \frac{1}{k}(U_{i,j+1} - U_{i,j}) - \frac{1}{h^2}(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}),$$

where $U_{i,j}$ is the value of the true solution U of the partial differential equation at the point (ih, jk). We have also implied that some formulas are locally more accurate than others, meaning that for the same h and k the local truncation errors $T_{i,j}$ are smaller in magnitude.

We now distinguish between what we mean by the *local* error and the *global* error at a point. We view *local* errors in the following way. Suppose we know the true solution of the differential equation at all points up to and including those at time level j. Then if we were to use, for example, the simple explicit scheme to find an approximation $u_{i,j+1}$ to $U_{i,j+1}$ we would calculate

$$u_{i,j+1} = U_{i,j} + r(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}),$$

where $r = k/h^2$.

By the definition of the local truncation error given above we see that

$$U_{i,j+1} - u_{i,j+1} = kT_{i,j}.$$

In this sense the $T_{i,j}$ measure the (local) error made in using the finite-difference scheme *once only*.

In a complete step-by-step process we apply the finite-difference scheme many times; the local errors accumulate and it is this accumulation which produces the global error. Thus, at any stage, we can define the *global error* as the difference $U_{i,j} - u_{i,j}$ between the computed finite-difference solution and the true solution of the differential equation at a point. It should be clear from the unstable example in Section 5.2 that simply reducing the local error does not ensure that the global error will be reduced; the local truncation error, however, will provide a measure of accuracy when there is no accumulation of errors.

We now show how to find an expression for the local truncation error $T_{i,j}$ for any finite-difference scheme applied to a differential equation. Obviously we cannot compute $T_{i,j}$ in general because we will not always know the values of the true solution of the differential equation. What we can do is to use Taylor's theorem to give us an expression for $T_{i,j}$ involving the mesh spacings h and k and certain derivatives of the true solution of the partial differential equation.

Again, in general, we will not know the values of the derivatives and at first sight we would appear to have gained no advantage. Fortunately, the expressions we obtain can be used to give a relative evaluation of various different schemes applied to a given differential equation.

Before looking at a specific example note that we can write a finite-difference scheme in operator notation as

$$Lu - b = 0$$

where L is a finite-difference operator which replaces some differential operator and b is known at each mesh point and depends on initial and/or boundary conditions. The local truncation error, $T_{i,j}$, is then the error which arises when the true solution of the differential equation is substituted on the left-hand side of the difference equation. Thus

$$T_{i,j} = LU_{i,j} - b_{i,j}.$$

To illustrate the complete method for finding $T_{i,j}$ let us consider the simple explicit scheme

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2}, \qquad (1)$$

applied to the partial differential equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2},$$

in which the operator $L = \Delta_t/k - \delta_x^2/h^2$ replaces

$$\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}.$$

We can write the local truncation error of scheme (1) as

$$T_{i,j} = \frac{1}{k}(U_{i,j+1} - U_{i,j}) - \frac{1}{h^2}(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}). \qquad (2)$$

Taylor's Theorem gives

$$U_{i,j+1} - U_{i,j} = k\frac{\partial U_{i,j}}{\partial t} + \frac{1}{2}k^2\frac{\partial^2 U_{i,j}}{\partial t^2} + O(k^3)$$

and

$$\begin{aligned} U_{i+1,j} - 2U_{i,j} + U_{i-1,j} &= (U_{i+1,j} - U_{i,j}) + (U_{i-1,j} - U_{i,j}) \\ &= h^2\frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{1}{12}h^4\frac{\partial^4 U_{i,j}}{\partial x^4} + O(h^6), \end{aligned} \qquad (3)$$

under suitable conditions. Therefore, we can write Equation (2) as

$$T_{i,j} = \frac{\partial U_{i,j}}{\partial t} - \frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{k}{2}\frac{\partial^2 U_{i,j}}{\partial t^2} - \frac{h^2}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(k^2) + O(h^4).$$

Since U satisfies the differential equation we have

$$\frac{\partial U_{i,j}}{\partial t} - \frac{\partial^2 U_{i,j}}{\partial x^2} = 0,$$

and so

$$T_{i,j} = \frac{k}{2}\frac{\partial^2 U_{i,j}}{\partial t^2} - \frac{h^2}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(k^2) + O(h^4) \qquad (4)$$

or, more simply,

$$T_{i,j} = O(k) + O(h^2).$$

Sometimes in the literature the local error is calculated from the form of the difference equation given by

$$u_{i,j+1} = u_{i,j} + r(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}).$$

That is, we would calculate

$$T^*_{i,j+1} = U_{i,j+1} - U_{i,j} - r(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}).$$

It is easy to see that $T^*_{i,j} = kT_{i,j+1}$. Either definition is acceptable and we shall use both in our numerical work. We have used the subscript $i, j + 1$ in our definition of T^* because this is the local truncation error made in calculating $u_{i,j+1}$.

SAQ 14

Show that the local truncation error of the Crank–Nicolson implicit method applied to the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}$$

is $O(k^2) + O(h^2)$.

(Solution on p. 41.)

SAQ 15

Find the local truncation error of the formula

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

when applied to the differential equation

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial^2 U}{\partial x^2}.$$

If the mesh ratio $p^2 = k^2/h^2 = 1$ what value does the local truncation error take?

(Solution on p. 42.)

SAQ 16

For what value of $r = k/h^2$ in the explicit scheme (1) would you obtain a much smaller local truncation error? (Use Equation (4) of this section as your starting point.)

(Solution on p. 43.)

5.6 SUMMARY

In this unit we have discussed and illustrated the following techniques:

how partial derivatives may be approximated by *finite-difference formulas* using Taylor approximations;

how to solve an initial-boundary value problem by replacing the differential equation by a finite-difference formula which is solved over a finite *mesh* of points in the xt-plane with corresponding initial and boundary conditions;

the representation of a finite-difference scheme by a *molecule*;

how to construct an *explicit* scheme for both the heat conduction and wave equations in one dimension;

the Crank–Nicolson *implicit* method which leads to a *tridiagonal* system of equations;

a direct method based on Gauss elimination, for the solution of a tridiagonal system of equations, which is conditionally stable and computationally efficient;

how to include derivative initial or boundary conditions in the numerical scheme;

the use of Taylor's Theorem to find the order of the *local truncation error* of a finite-difference scheme.

We have also discussed induced instability in a numerical scheme due to the accumulation of local errors. For the explicit scheme this depends on the value of the *mesh ratio*. We have seen that a scheme determines the *numerical domain of influence* of a point and that the *finite-difference replacement* of a parabolic differential equation may have two families of *numerical characteristics*.

5.7 FURTHER SELF-ASSESSMENT QUESTIONS

If you are short of time you may omit this section and return to it for revision purposes.

SAQ 17

S: page 46, Exercise 2
You need not derive the analytical solution.

(Solution on p. 43.)

You may find the next SAQ somewhat tricky.

SAQ 18

Suggest an explicit finite-difference scheme for the diffusion equation in two dimensions,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Use the same mesh spacing h in both the x- and y-directions. Draw the molecular diagram for the scheme.

HINT: Draw a diagram (three-dimensional since u is a function of three variables) before devising the finite-difference scheme.

As far as the actual computation is concerned, what added complications arise in this problem compared with the equation in one dimension,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}?$$

(Solution on p. 43.)

SAQ 19

Describe a numerical experiment to solve the initial-value problem with which this unit began:

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) + (\sin xt)u(x, t) = 0 \qquad 0 < x < 1, \quad t > 0,$$

$$u(0, t) = u(1, t) = 0 \qquad t \geqslant 0,$$

$$u(x, 0) = f(x) \qquad 0 \leqslant x \leqslant 1.$$

(Solution on p. 44.)

5.8 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

Using the central-difference formulas of Equations (1.8) and (1.9) in *S: page 8* we obtain

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

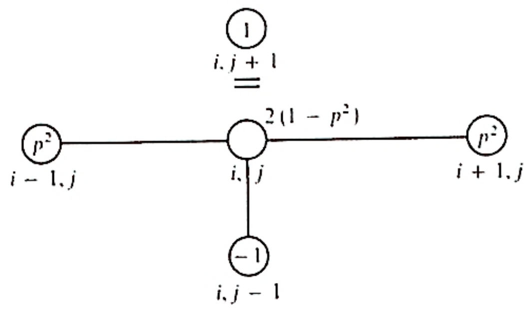
as the finite-difference replacement of

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

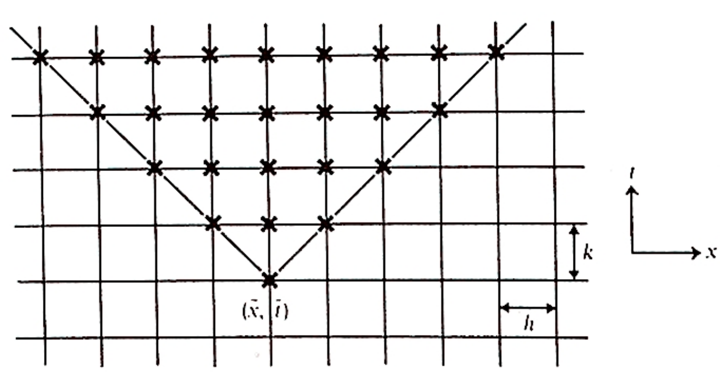
We can write the finite-difference scheme as

$$u_{i,j+1} = 2(1 - p^2)u_{i,j} + p^2(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1}$$

where $p^2 = k^2/h^2$ is the *mesh ratio*. The scheme, having only one unknown $u_{i,j+1}$, is explicit. The molecule for the scheme is shown in the diagram.



For the numerical domain of influence we show all those points which depend for their value of u on the value at $(\bar{x}, \bar{t})$ by crosses in the diagram.



The numerical characteristics through $(\bar{x}, \bar{t})$ are

$$t - \bar{t} = \pm p(x - \bar{x}),$$

and the numerical domain of influence is given by

$$\{(\bar{x} + ih, \bar{t} + jk) : i = -j, -j+1, \ldots, j-1, j; j = 0, 1, 2, \ldots\}.$$

Since the scheme uses information on two levels (j and $j - 1$) we need to be able to specify values along $t = 0$ and $t = k$ before the scheme can be used. For the Cauchy problem these are available since both u and $\partial u/\partial t$ are specified for $t = 0$. We shall see how this works in practice in Section 5.4.

Solution to SAQ 2

(a) 50 LET U[I, J + 1] = U[I, J] + R * (U[I − 1, J] − 2 * U[I, J] + U[I + 1, J])

where U is a two-dimensional array which has been declared earlier in the program to be big enough to hold the solution to the differential equation at all points in the solution domain.

(Generally speaking we would only require to store values of u along two time levels, j and $j+1$, since once we have calculated u at all points along level $j+1$ the values along level j are no longer required. We could then replace U[I, J] by U[I, J + 1] and repeat the calculations to yield

$$u_{i,j+2} \qquad i = 1, 2, \ldots, N.)$$

(b)
```
45  FOR I = 2 TO N - 1
50  LET U[I, J + 1] = U[I, J] + R * (U[I - 1, J] - 2 * U[I, J] + U[I + 1, J])
55  NEXT I
```

(c)
```
40  FOR J = 1 TO M
45  FOR I = 2 TO N - 1
50  LET U[I, J + 1] = U[I, J] + R * (U[I - 1, J] - 2 * U[I, J] + U[I + 1, J])
55  NEXT I
60  NEXT J
```

(d) (i) $u_{1,j}$ and $u_{N,j}$ ($j = 1, 2, \ldots, M$) are the boundary values.
(ii) $u_{i,1}$ ($i = 1, 2, \ldots, N$) are the initial values.

These values are assumed to be given so that they may be read in to the computer as data.

(e) In BASIC we are not allowed to use the zero subscript. We have therefore assumed that the left-hand boundary is at $i = 1$ and that the initial data are given along $j = 1$.

(f)

```
10   PRINT "THIS PROGRAM EVALUATES THE HEAT CONDUCTION EQUATION IN ONE"
20   PRINT "SPACE VARIABLE USING THE SIMPLE EXPLICIT FINITE DIFFERENCE"
30   PRINT "SCHEME"
40   DIM U[20,21]
50   PRINT "TYPE NUMBER OF POINTS IN X-DIRECTION AND SIZE OF TIME STEP"
60   INPUT N,K
70   H=3.14159/(N-1)
80   R=K/(H*H)
90   M=20
100  L=INT((N+1)/2)
110  X=(L-1)*H
120  PRINT "VALUE OF R IS";R
130  PRINT
140  PRINT "                    SOLUTION ALONG X=";X
150  PRINT "                    ----------------------------"
160  PRINT "TIME","THEOR.VAL","CALC.VAL.","ERROR"
170  PRINT
180  FOR I=2 TO N-1
190  U[I,1]=SIN((I-1)*H)
200  NEXT I
210  PRINT 0,SIN((L-1)*H),U[L,1],0
220  PRINT
230  FOR J=1 TO M
240  U[1,J]=U[N,J]=0
250  NEXT J
260  FOR J=1 TO M
270  T=J*K
280  FOR I=2 TO N-1
290  U[I,J+1]=U[I,J]+R*(U[I+1,J]-2*U[I,J]+U[I-1,J])
300  NEXT I
310  A=EXP(-T)*SIN(X)
320  C=U[L,J+1]
330  PRINT T,A,C,A-C
340  PRINT
350  NEXT J
360  END
```

The preceding listing is the program $EXP321.

(g) The following results are a typical set, obtained with $N = 5$, $k = 0.1$.

```
RUN
EXP321

THIS PROGRAM EVALUATES THE HEAT CONDUCTION EQUATION IN ONE
SPACE VARIABLE USING THE SIMPLE EXPLICIT FINITE DIFFERENCE
SCHEME
TYPE NUMBER OF POINTS IN X-DIRECTION AND SIZE OF TIME STEP?5,Ø.1
VALUE OF R IS .162114

                SOLUTION ALONG X= 1.57Ø8
                ----------------------------
```

TIME	THEOR.VAL	CALC.VAL.	ERROR
Ø	1.	1.	Ø
.1	.9Ø4838	.9Ø5Ø36	-1.98245E-Ø4
.2	.818731	.819Ø9	-3.58939E-Ø4
.3	.74Ø818	.7413Ø6	-4.87328E-Ø4
.4	.67Ø32	.67Ø9Ø8	-5.8794ØE-Ø4
.5	.6Ø6531	.6Ø7196	-6.65188E-Ø4
.6	.548812	.549534	-7.224Ø3E-Ø4
.7	.496585	.497348	-7.62463E-Ø4
.8	.449329	.45Ø118	-7.8845ØE-Ø4
.9	.4Ø657	.4Ø7373	-8.Ø2755E-Ø4
1	.367879	.368687	-8.Ø7285E-Ø4
1.1	.332871	.333675	-8.Ø3471E-Ø4
1.2	.3Ø1194	.3Ø1987	-7.93219E-Ø4
1.3	.272532	.2733Ø9	-7.776Ø2E-Ø4
1.4	.246597	.247355	-7.57813E-Ø4
1.5	.22313	.223865	-7.34776E-Ø4
1.6	.2Ø1897	.2Ø26Ø6	-7.Ø9146E-Ø4
1.7	.182684	.183365	-6.81877E-Ø4
1.8	.165299	.165952	-6.53386E-Ø4
1.9	.149569	.15Ø193	-6.24Ø9ØE-Ø4
2	.135335	.13593	-5.94527E-Ø4

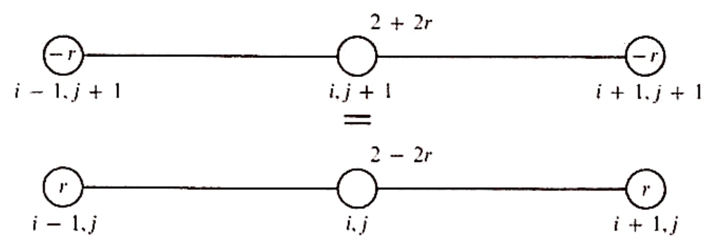
DONE

Solution to SAQ 3

(a) The Crank–Nicolson method, given in *S: page 17*, is

$$-ru_{i-1,j+1}+(2+2r)u_{i,j+1}-ru_{i+1,j+1}=ru_{i-1,j}+(2-2r)u_{i,j}+ru_{i+1,j}.$$

Its molecule is



(b) We have

$$\delta_x^2 u_{i,j} = u_{i+1,j} - 2u_{i,j} + u_{i-1,j}$$

and

$$\delta_x^2 u_{i,j+1} = u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}.$$

Therefore, we can write the Crank–Nicolson method as

$$2u_{i,j+1} - r\delta_x^2 u_{i,j+1} = r\delta_x^2 u_{i,j} + 2u_{i,j},$$

or

$$(2 - r\delta_x^2)u_{i,j+1} = (2 + r\delta_x^2)u_{i,j}.$$

Solution to SAQ 4

Let the two-dimensional array U[I, J] (I = 1, 2, ..., N; J = 1, 2, ..., M) hold the values of the solution of the finite-difference scheme at the mesh points $(x, t) = (ih, jk)$. The coefficients of the unknowns at the $(j + 1)$th time level are to be stored in the $(N - 2) \times (N - 2)$ array C.

We assume that the arrays C and U have already been declared and that the solution to the differential equation is known at all time levels up to and including $t = jk$.

```
50 FOR L = 2 TO N - 3
60 LET C[L, L - 1] = -R
61 LET C[L, L] = 2 + 2 * R
62 LET C[L, L + 1] = -R
70 NEXT L
80 LET C[1, 1] = 2 + 2 * R
81 LET C[1, 2] = -R
85 LET C[N - 2, N - 3] = -R
86 LET C[N - 2, N - 2] = 2 + 2 * R
```

The boundary values, which are known, are assumed to occupy U[1, J] and U[N, J] (J = 1, 2, ..., M). As a result there are $N - 2$ unknowns to be found at each time level.

(Note that the piece of program in this solution is inefficient. In practice we would first calculate $2 + 2r$ and $-r$ and use stored values for the LET assignments.)

Solution to SAQ 5

The Crank–Nicolson formula is

$$-ru_{i-1,j+1} + (2 + 2r)u_{i,j+1} - ru_{i+1,j+1} = ru_{i-1,j} + (2 - 2r)u_{i,j} + ru_{i+1,j}.$$

For six internal mesh points let $i = 0, 1, 2, \ldots, 7$ where $i = 0$ and $i = 7$ are the boundary points. The relevant equations are

$$-ru_{0,j+1} + (2 + 2r)u_{1,j+1} - ru_{2,j+1} = b_1$$
$$-ru_{1,j+1} + (2 + 2r)u_{2,j+1} - ru_{3,j+1} = b_2$$
$$-ru_{2,j+1} + (2 + 2r)u_{3,j+1} - ru_{4,j+1} = b_3$$
$$-ru_{3,j+1} + (2 + 2r)u_{4,j+1} - ru_{5,j+1} = b_4$$
$$-ru_{4,j+1} + (2 + 2r)u_{5,j+1} - ru_{6,j+1} = b_5$$
$$-ru_{5,j+1} + (2 + 2r)u_{6,j+1} - ru_{7,j+1} = b_6$$

where b_1, b_2, b_3, b_4, b_5 and b_6 are obtained from the known values of u along the jth time level. Hence, in matrix form,

$$\begin{bmatrix} 2+2r & -r & & & & \\ -r & 2+2r & -r & & & \\ & -r & 2+2r & -r & & \\ & & -r & 2+2r & -r & \\ & & & -r & 2+2r & -r \\ & & & & -r & 2+2r \end{bmatrix} \begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ u_{4,j+1} \\ u_{5,j+1} \\ u_{6,j+1} \end{bmatrix} = \begin{bmatrix} b_1 + ru_{0,j+1} \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 + ru_{7,j+1} \end{bmatrix}$$

where $u_{0,j+1}$ and $u_{7,j+1}$ are known boundary values.

The matrix is tridiagonal and also symmetric.

Solution to SAQ 6

At $(x, t) = (0.5, 0.1)$ the solution of the problem using the Crank–Nicolson implicit scheme with $r = 0.5$ and $h = 0.1$ is 0.3756, with a percentage error of 0.8. This is better than the explicit method with $r = 0.5$ but not so good as the explicit method with $r = 0.1$. (Further results are presented in ***S:*** *pages 49 and 50.*)

Solution to SAQ 7

Putting $\theta = 1$ in the weighted average approximation gives the scheme

$$u_{i,j+1} - u_{i,j} = r(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1})$$

which can be written as

$$-ru_{i+1,j+1} + (2r + 1)u_{i,j+1} - ru_{i-1,j+1} = u_{i,j}.$$

The values of u for the first time step satisfy

$$\begin{aligned} 3u_1 - u_2 &= 0.2, \\ -u_1 + 3u_2 - u_3 &= 0.4, \\ -u_2 + 3u_3 - u_4 &= 0.6, \\ -u_3 + 3u_4 - u_5 &= 0.8, \\ -2u_4 + 3u_5 &= 1.0, \end{aligned}$$

where u_i denotes $u_{i,1}$. To obtain the last equation we have taken advantage of the symmetry of the problem. Using the notation of *S: page 20* we have

$$a_2 = a_3 = a_4 = 1, a_5 = 2,$$

$$b_1 = b_2 = b_3 = b_4 = b_5 = 3,$$

$$c_1 = c_2 = c_3 = c_4 = 1,$$

$$d_1 = 0.2, d_2 = 0.4, d_3 = 0.6, d_4 = 0.8, d_5 = 1.0.$$

We now define α_i and S_i recursively, for $i = 1, 2, 3, 4, 5$, using the equations at the top of *S: page 22*.

$$\alpha_1 = 3 \qquad S_1 = \tfrac{1}{5}$$

$$\alpha_2 = 3 - \tfrac{1}{3} = \tfrac{8}{3} \qquad S_2 = \tfrac{2}{5} + \tfrac{1}{3}\cdot\tfrac{1}{5} = \tfrac{7}{15}$$

$$\alpha_3 = 3 - \tfrac{3}{8} = \tfrac{21}{8} \qquad S_3 = \tfrac{3}{5} + \tfrac{3}{8}\cdot\tfrac{7}{15} = \tfrac{31}{40}$$

$$\alpha_4 = 3 - \tfrac{8}{21} = \tfrac{55}{21} \qquad S_4 = \tfrac{4}{5} + \tfrac{8}{21}\cdot\tfrac{31}{40} = \tfrac{23}{21}$$

$$\alpha_5 = 3 - \tfrac{2.21}{55} = \tfrac{123}{55} \qquad S_5 = 1 + \tfrac{2.21}{55}\cdot\tfrac{23}{21} = \tfrac{101}{55}.$$

Therefore, applying the recurrence relation

$$u_5 = \frac{S_5}{\alpha_5},$$

$$u_i = \frac{1}{\alpha_i}(S_i + c_i u_{i+1}),$$

we obtain

$$u_5 = \tfrac{101}{123} \simeq 0.821,$$

$$u_4 = \tfrac{30}{41} \simeq 0.732,$$

$$u_3 = \tfrac{353}{615} \simeq 0.574,$$

$$u_2 = \tfrac{16}{41} \simeq 0.390,$$

$$u_1 = \tfrac{121}{615} \simeq 0.197.$$

Solution to SAQ 8

The following is a print-out of the library program STR1321 which uses the recurrence method for the solution of a tridiagonal system of linear equations.

```
10   PRINT "PROGRAM TO SOLVE A SYSTEM OF TRIDIAGONAL LINEAR EQUATIONS"
20   PRINT
30   REM  ***  THE INPUT STAGE  ***
40   PRINT "TYPE THE NUMBER OF EQUATIONS(<=10)";
50   INPUT N
60   PRINT "TYPE 3 SIGNIFICANT ELEMENTS OF TRIDIAGONALS FOR EACH ROW"
70   PRINT "    (N.B. ONLY 2 ELEMENTS FOR FIRST AND LAST ROWS)"
80   PRINT "FIRST ROW";
90   INPUT B[1],C[1]
100   FOR I=2 TO N-1
110   PRINT "ROW";I;
120   INPUT A[I],B[I],C[I]
130   NEXT I
140   PRINT "LAST ROW ";
150   INPUT A[N],B[N]
160   PRINT "TYPE R.H.S. CONSTANTS"
170   MAT  INPUT D[N]
180   REM ***   CALCULATE THE L(I) AND S(I)   ***
190   S[1]=D[1]
```

```
200  L[1]=B[1]
210  FOR I=2 TO N
220  Z=A[I]/L[I-1]
230  L[I]=B[I]-C[I-1]*Z
240  S[I]=D[I]-S[I-1]*Z
250  NEXT I
260  REM  ***  CALCULATE THE U(I)  ***
270  U[N]=S[N]/L[N]
280  FOR I=N-1 TO 1 STEP -1
290  U[I]=(S[I]-C[I]*U[I+1])/L[I]
300  NEXT I
310  REM  ***  OUTPUT STAGE  ***
320  PRINT
330  PRINT "SOLUTION"
340  PRINT "--------"
350  FOR I=1 TO N
360  PRINT U[I]
370  NEXT I
380  END
```

Note that the arrays store the actual coefficients, so that if a_i and c_i are defined as in *S: page 20* then A[I] and C[I] store the values $-a_i$ and $-c_i$.

Solution to SAQ 9

(i) Using the notation of *S: page 20* we obtain

$$a_2 = a_3 = a_4 = \frac{3}{2},$$

$$b_1 = b_2 = b_3 = b_4 = 2,$$

$$c_1 = c_2 = c_3 = \frac{3}{2},$$

$$d_1 = -1, d_2 = -2, d_3 = -3, d_4 = \frac{7}{2}.$$

Hence

$$\alpha_1 = b_1 = 2,$$

$$\alpha_2 = b_2 - \frac{a_2}{\alpha_1}c_1 = 2 - \frac{3}{2}\cdot\frac{1}{2}\cdot\frac{3}{2} = \frac{7}{8},$$

$$\alpha_3 = b_3 - \frac{a_3}{\alpha_2}c_2 = 2 - \frac{3}{2}\cdot\frac{8}{7}\cdot\frac{3}{2} = -\frac{4}{7},$$

$$\alpha_4 = b_4 - \frac{a_4}{\alpha_3}c_3 = 2 + \frac{3}{2}\cdot\frac{7}{4}\cdot\frac{3}{2} = \frac{95}{16},$$

and

$$S_1 = d_1 = -1,$$

$$S_2 = d_2 + \frac{a_2}{\alpha_1}S_1 = -2 - \frac{3}{2}\cdot\frac{1}{2} = -\frac{11}{4},$$

$$S_3 = d_3 + \frac{a_3}{\alpha_2}S_2 = -3 - \frac{3}{2}\cdot\frac{8}{7}\cdot\frac{11}{4} = -\frac{54}{7},$$

$$S_4 = d_4 + \frac{a_4}{\alpha_3}S_3 = \frac{7}{2} + \frac{3}{2}\cdot\frac{7}{4}\cdot\frac{54}{7} = \frac{95}{4}.$$

Therefore

$$x_4 = \frac{S_4}{\alpha_4} = \frac{95}{4}\cdot\frac{16}{95} = 4,$$

$$x_3 = \frac{1}{\alpha_3}(S_3 + c_3x_4) = -\frac{7}{4}\left(-\frac{54}{7} + \frac{3}{2}\cdot 4\right) = 3,$$

$$x_2 = \frac{1}{\alpha_2}(S_2 + c_2x_3) = \frac{8}{7}\left(-\frac{11}{4} + \frac{3}{2}\cdot 3\right) = 2,$$

$$x_1 = \frac{1}{\alpha_1}(S_1 + c_1x_2) = \frac{1}{2}\left(-1 + \frac{3}{2}\cdot 2\right) = 1.$$

(ii) The output from the program $TRI321 for this problem follows.

```
RUN
TRI321

PROGRAM TO SOLVE A SYSTEM OF TRIDIAGONAL LINEAR EQUATIONS

TYPE THE NUMBER OF EQUATIONS(<=10)?4
TYPE 3 SIGNIFICANT ELEMENTS OF TRIDIAGONALS FOR EACH ROW
     (N.B. ONLY 2 ELEMENTS FOR FIRST AND LAST ROWS)
FIRST ROW?2,-1.5
ROW 2     ?-1.5,2,-1.5
ROW 3     ?-1.5,2,-1.5
LAST ROW ?-1.5,2
TYPE R.H.S. CONSTANTS
?-1,-2,-3,3.5

SOLUTION
--------
 1.
 2.
 3.
 4

DONE
```

Solution to SAQ 10

The implicit method for the hyperbolic equation is

$$\delta_t^2 u_{i,j} = p^2\{\theta\delta_x^2 u_{i,j+1} + (1 - 2\theta)\delta_x^2 u_{i,j} + \theta\delta_x^2 u_{i,j-1}\}$$

where $p^2 = k^2/h^2$. Assuming that the values along time levels j and $j - 1$ are known this gives rise to the equation

$$u_{i,j+1} = p^2\theta(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}) + d_i,$$

where d_i depends only on known values. If we assume that the boundary values $u_{0,j+1}$ and $u_{N,j+1}$ are also known, then, combining the equations for $i = 1, \ldots, N - 1$ we obtain

$$\begin{bmatrix} 2p^2\theta + 1 & -p^2\theta & & & \\ -p^2\theta & 2p^2\theta + 1 & -p^2\theta & & \\ & -p^2\theta & 2p^2\theta + 1 & -p^2\theta & \\ & & \cdot & & \cdot \\ & & & -p^2\theta & 2p^2\theta + 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \cdot \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} d_1 + p^2\theta u_0 \\ d_2 \\ d_3 \\ \cdot \\ d_{N-1} + p^2\theta u_N \end{bmatrix}$$

where we now write u_i for $u_{i,j+1}$. It is clear that, in the notation of *S: page 20*, $a_i > 0$, $b_i > 0$, $c_i > 0$, $b_i > a_i + c_i$ and $b_i > a_{i+1} + c_{i-1}$. Therefore, the recurrence

method for solving tridiagonal systems (which is equivalent to Gauss elimination without pivoting) is stable when applied to this system.

Solution to SAQ 11

We saw in **S**: *pages 6 and 7* that the central-difference approximation to a first derivative has an error of order h^2 whereas the error in the forward-difference approximation is of order h. Since h is less than unity, the conclusion follows.

Solution to SAQ 12

We use the explicit scheme

$$u_{i,j+1} = u_{i,j} + r(u_{i-1,j} - 2u_{i,j} + u_{i+1,j})$$

(Equation (2.4) in **S**: *page 10*).

(a) For the internal mesh points ($i = 1, 2, \ldots, N-1$) the explicit scheme may be used directly. At the boundary $x = 0$ the condition

$$\frac{\partial u}{\partial x} = h_1(u - v_1)$$

may be approximated by the central-difference formula

$$\frac{1}{2\delta x}(u_{1,j} - u_{-1,j}) = h_1(u_{0,j} - v_1).$$

Elimination of $u_{-1,j}$ between this equation and

$$u_{0,j+1} = u_{0,j} + r(u_{-1,j} - 2u_{0,j} + u_{1,j})$$

yields

$$u_{0,j+1} = \{1 - 2r(1 + h_1\delta x)\}u_{0,j} + 2ru_{1,j} + 2rh_1v_1\delta x.$$

Similar treatment of the boundary $x = 1$ yields

$$u_{N,j+1} = 2ru_{N-1,j} + \{1 - 2r(1 + h_2\delta x)\}u_{N,j} + 2rh_2v_2\delta x.$$

(b) At the boundary $x = 0$ the condition

$$\frac{\partial u}{\partial x} = h_1(u - v_1)$$

may be approximated by the forward-difference formula

$$\frac{1}{\delta x}(u_{1,j} - u_{0,j}) = h_1(u_{0,j} - v_1).$$

The scheme yields

$$\begin{aligned} u_{1,j+1} &= u_{1,j} + r(u_{0,j} - 2u_{1,j} + u_{2,j}) \\ &= \{1 - 2r + r/(1 + h_1\delta x)\}u_{1,j} + ru_{2,j} + rh_1v_1\delta x/(1 + h_1\delta x), \end{aligned}$$

and the forward-difference formula may be applied at level $j + 1$ to give

$$u_{0,j+1} = (u_{1,j+1} + h_1v_1\delta x)/(1 + h_1\delta x).$$

The equations for $u_{N-1,j+1}$ and $u_{N,j+1}$ are obtained similarly.

Solution to SAQ 13

Modelling the problem is straightforward. The solution to the problem appears in **S**: *pages 51 to 53*. You need not verify the analytical solution.

The function erfc in the solution arises as follows. Many functions have indefinite integrals which cannot be expressed by the use of a finite number of elementary functions (polynomials, sin, cos, exp, log). Many of these integrals occur so frequently

that they have been given names and have been tabulated as new functions. One such function is the (so called) **error function** defined by

$$\operatorname{erf}(x) = \frac{2}{\pi^{1/2}}\int_0^x e^{-\xi^2}d\xi \qquad x \in R.$$

The **complementary error function** is defined by

$$\operatorname{erfc}(x) = \frac{2}{\pi^{1/2}}\int_x^\infty e^{-\xi^2}d\xi \qquad x \in R,$$

from which it can be shown that

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \qquad x \in R.$$

Values of the functions erf and erfc are tabulated in, for example, M. Abramowitz and I. Stegun (eds.), *Handbook of Mathematical Functions* (Dover, 1965).

Solution to SAQ 14

The Crank–Nicolson formula is given on *S: page 17*, and the local truncation error is

$$T_{i,j} = \frac{1}{k}(U_{i,j+1} - U_{i,j}) - \frac{1}{2h^2}(\delta_x^2 U_{i,j+1} + \delta_x^2 U_{i,j}),$$

where U satisfies

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}.$$

Now, by Taylor's Theorem,

$$U_{i,j+1} = U_{i,j} + k\frac{\partial U_{i,j}}{\partial t} + \frac{k^2}{2}\frac{\partial^2 U_{i,j}}{\partial t^2} + O(k^3)$$

under suitable conditions.

As in Equation (3) of Section 5.5 we have

$$\delta_x^2 U_{i,j} = h^2\frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(h^6),$$

and

$$\delta_x^2 U_{i,j+1} = h^2\frac{\partial^2 U_{i,j+1}}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 U_{i,j+1}}{\partial x^4} + O(h^6).$$

Applying Taylor's Theorem to $\partial^2 U/\partial x^2$ and $\partial^4 U/\partial x^4$ we obtain

$$\frac{\partial^2 U_{i,j+1}}{\partial x^2} = \frac{\partial^2 U_{i,j}}{\partial x^2} + k\frac{\partial^3 U_{i,j}}{\partial t\partial x^2} + O(k^2)$$

and

$$\frac{\partial^4 U_{i,j+1}}{\partial x^4} = \frac{\partial^4 U_{i,j}}{\partial x^4} + O(k).$$

Thus

$$T_{i,j} = \left[\frac{\partial U_{i,j}}{\partial t} - \frac{\partial^2 U_{i,j}}{\partial x^2}\right] + \frac{k}{2}\frac{\partial}{\partial t}\left[\frac{\partial U_{i,j}}{\partial t} - \frac{\partial^2 U_{i,j}}{\partial x^2}\right] - \frac{h^2}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(k^2) + O(h^4) + O(kh^2).$$

Since U satisfies the differential equation,

$$\frac{\partial U_{i,j}}{\partial t} = \frac{\partial^2 U_{i,j}}{\partial x^2},$$

and the local truncation error is

$$T_{i,j} = \frac{-h^2}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(k^2) + O(h^4) + O(kh^2)$$

$$= O(k^2) + O(h^2).$$

Solution to SAQ 15

Expansion in Taylor series yields

$$U_{i,j+1} - 2U_{i,j} + U_{i,j-1} = k^2 \frac{\partial^2 U_{i,j}}{\partial t^2} + \frac{k^4}{12}\frac{\partial^4 U_{i,j}}{\partial t^4} + \frac{k^6}{360}\frac{\partial^6 U_{i,j}}{\partial t^6} + \cdots$$

and

$$U_{i+1,j} - 2U_{i,j} + U_{i-1,j} = h^2 \frac{\partial^2 U_{i,j}}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + \frac{h^6}{360}\frac{\partial^6 U_{i,j}}{\partial x^6} + \cdots.$$

Therefore,

$$T_{i,j} = \left(\frac{\partial^2 U_{i,j}}{\partial t^2} - \frac{\partial^2 U_{i,j}}{\partial x^2}\right) + \frac{1}{12}\left(k^2 \frac{\partial^4 U_{i,j}}{\partial t^4} - h^2 \frac{\partial^4 U_{i,j}}{\partial x^4}\right)$$
$$+ \frac{1}{360}\left(k^4 \frac{\partial^6 U_{i,j}}{\partial t^6} - h^4 \frac{\partial^6 U_{i,j}}{\partial x^6}\right) + \cdots.$$

The differential equation gives

$$\frac{\partial^2 U}{\partial t^2} - \frac{\partial^2 U}{\partial x^2} = 0,$$

so that

$$T_{i,j} = \frac{1}{12}\left(k^2 \frac{\partial^4 U_{i,j}}{\partial t^4} - h^2 \frac{\partial^4 U_{i,j}}{\partial x^4}\right) + \frac{1}{360}\left(k^4 \frac{\partial^6 U_{i,j}}{\partial t^6} - h^4 \frac{\partial^6 U_{i,j}}{\partial x^6}\right) + \cdots$$

and the local truncation error of the scheme is $O(k^2) + O(h^2)$.

If $p = 1$ then $k = h$ and

$$T_{i,j} = \frac{h^2}{12}\left(\frac{\partial^4 U_{i,j}}{\partial t^4} - \frac{\partial^4 U_{i,j}}{\partial x^4}\right) + \frac{h^4}{360}\left(\frac{\partial^6 U_{i,j}}{\partial t^6} - \frac{\partial^6 U_{i,j}}{\partial x^6}\right) + \cdots.$$

Now, assuming that U has partial derivatives of all orders,

$$\begin{aligned}
\frac{\partial^4 U}{\partial t^4} &= \frac{\partial^2}{\partial t^2}\left(\frac{\partial^2 U}{\partial x^2}\right) && \text{using the differential equation}\\
&= \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 U}{\partial t^2}\right) && \text{since all derivatives exist}\\
&= \frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 U}{\partial x^2}\right) && \text{using the differential equation}\\
&= \frac{\partial^4 U}{\partial x^4}.
\end{aligned}$$

We may show, by induction, that

$$\frac{\partial^{2n} U}{\partial t^{2n}} = \frac{\partial^{2n} U}{\partial x^{2n}} \qquad n = 1, 2, \ldots$$

whence $T_{i,j} = 0$. We have shown that the finite-difference scheme

$$u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}$$

is an *exact* finite-difference representation of the wave equation.

Solution to SAQ 16

The local truncation error, given by Equation (4) of Section 5.5, is

$$T_{i,j} = \frac{k}{2}\frac{\partial^2 U_{i,j}}{\partial t^2} - \frac{h^2}{12}\frac{\partial^4 U_{i,j}}{\partial x^4} + O(k^2) + O(h^4).$$

Provided that the third derivatives of U all exist and are continuous at i,j we have, using the differential equation,

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial^3 U}{\partial t \partial x^2} = \frac{\partial^3 U}{\partial x^2 \partial t} = \frac{\partial^4 U}{\partial x^4} \qquad \text{at } i,j,$$

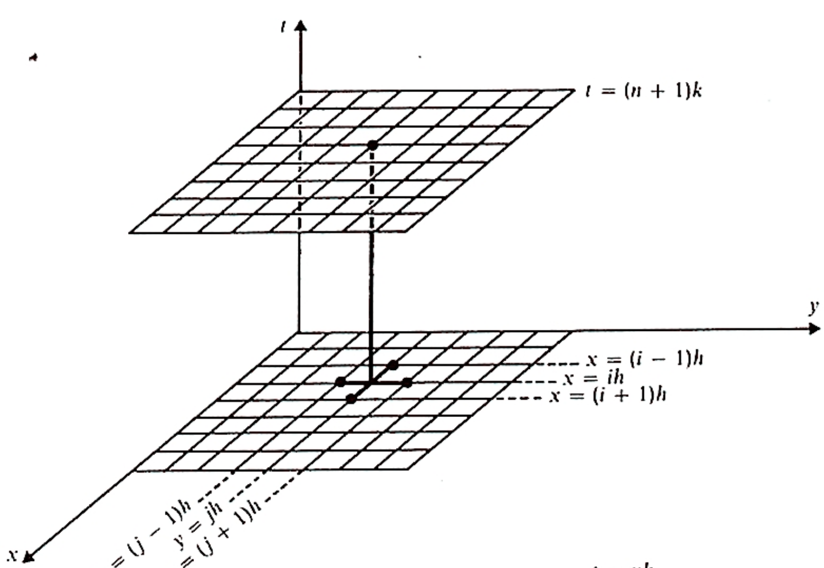
and we can reduce the error to $O(k^2) + O(h^4)$ by choosing $r = k/h^2 = \frac{1}{6}$.

(This is not very useful, however, since if h is small then $k = \frac{1}{6}h^2$ is very small and the resulting volume of computation is quite substantial.)

Solution to SAQ 17

See *S: pages 48 and 49.*
When it is not possible to remove a discontinuity by finding a transformation as suggested in *S*, it may be feasible to use very small values of the mesh spacings for the first few steps to improve the accuracy until the effect of the discontinuity has been sufficiently reduced to enable larger, more economic values to be used.

Solution to SAQ 18



The diagram shows two time levels, $t = nk$ and $t = (n+1)k$, where k is the mesh spacing in the t-direction. We replace the space derivatives by central differences and the time derivative by a forward difference. Using the notation

$$u_{i,j,n}$$

to represent $u(ih, jh, nk)$, we obtain

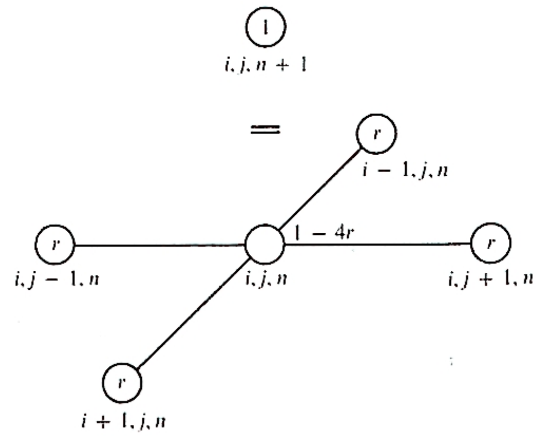
$$\frac{u_{i,j,n+1} - u_{i,j,n}}{k} = \frac{u_{i,j+1,n} - 2u_{i,j,n} + u_{i,j-1,n}}{h^2} + \frac{u_{i+1,j,n} - 2u_{i,j,n} + u_{i-1,j,n}}{h^2},$$

which can be written as

$$u_{i,j,n+1} = u_{i,j,n} + r(u_{i,j+1,n} + u_{i,j-1,n} + u_{i+1,j,n} + u_{i-1,j,n} - 4u_{i,j,n})$$

where $r = k/h^2$ is the mesh ratio. Since there is only one unknown at the $(n+1)$th time level the scheme is an explicit one.

The molecule of this scheme is shown in the diagram.



There is no difference, as far as the computations are concerned, between this case and the problem in one space dimension. There is, however, a greater number of mesh points to be taken into account. Note also that, provided we take a square mesh in the space dimensions, the mesh ratio is the same in both cases.

If we were to consider implicit methods then we should have to solve $(N - 1)^2$ equations in $(N - 1)^2$ unknowns at each time level (assuming a square mesh with $N - 1$ internal mesh points in both the x- and y-directions). The problem, for higher dimensions, is finding computationally faster methods for solving the much greater number of simultaneous linear equations which arise in the case of implicit methods.

The Crank–Nicolson implicit method gives

$$u_{i,j,n+1} - u_{i,j,n} = \frac{r}{2}\{\delta_x^2 u_{i,j,n+1} + \delta_y^2 u_{i,j,n+1} + \delta_x^2 u_{i,j,n} + \delta_y^2 u_{i,j,n}\}.$$

Solution to SAQ 19

We can write down the explicit numerical scheme

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} - \sin(ihjk)u_{i,j}$$

or the Crank–Nicolson implicit scheme

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{1}{2}\left\{\frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{h^2} + \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}\right\}$$
$$- \tfrac{1}{2}\{\sin(ih(j+1)k)u_{i,j+1} + \sin(ihjk)u_{i,j}\},$$

both of which can be solved by the techniques covered in this unit. The initial condition becomes

$$u_{i,0} = f(ih).$$

The only problem is the possibility of a discontinuity at $(0, 0)$ and $(1, 0)$ if $f(0)$ or $f(1)$ is nonzero. The effect of such a discontinuity can be reduced by taking small mesh spacings for the first time steps.

Unit 6 Fourier Series

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differential Equations* (Oxford, 1971).

H. F. Weinberger, *A First Course in Partial Differential Equations* (Blaisdell, 1965).

It is essential to have these books: the course is based on them and will not make sense without them. They are referred to in the text as S and W respectively.

Unit 6 is based on W: Chapter IV, Sections 14 to 16, 18, 20 to 23.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics.* References to Open University courses in mathematics take the form:

Unit M100 13, Integration II for the Mathematics Foundation Course,
Unit M201 23, The Wave Equation for the Linear Mathematics Course.

Bibliography

E. M. Horsburgh (Ed.), *Napier Tercentenary Celebration* (Royal Society, 1914).

This book was produced on the tercentenary of Napier's birth. Chapter 6 gives a good description of various harmonic analysers, describing both the mechanics and the mathematics.

J. Harvey, 'A Harmonic Analyser' in *Proceedings of the Physical Society*, 1930, Vol. 42, pp. 245–49.

This is a paper describing a mechanical device for analysing a function into its harmonics.

R. Furth and P. W. Pringle, 'A New Photo-Electric Method for Fourier Synthesis and Analysis' in *Philosophical Magazine*, 1944, Vol. 35, pp. 643–56.

This paper described an electronic harmonic analyser.

M. Spivak, *Calculus* (Benjamin, 1967).

This is the set book for the Analysis course, M231.

J. Fourier, *The Analytical Theory of Heat* (Dover, 1955).

This translation of Fourier's original work is primarily of historical interest.

6.0 INTRODUCTION

Our main aim in this unit is to revise one of the most common analytical methods of solving partial differential equations. The basic ideas have been covered before, in particular in *Unit M201 22, Fourier Series* and *Unit M201 32, The Heat Conduction Equation*, although a few of the concepts are new.

The first part of the unit (Section 6.1) deals with the method of separation of variables. Separability is an extremely important property of many linear partial differential equations, as it allows them to be transformed to systems of ordinary differential equations. In general, analytical solutions are available for separable problems only.

The ordinary differential equations that are obtained from separating the variables lead us to look for solutions which can be expressed as Fourier series.

Thus the next part of this unit comprises a revision of Fourier series, which were introduced in *Unit M201 22*, and a short study of their convergence properties; to do this it is necessary to introduce a few notions of real analysis, in particular the idea of *uniform convergence*.

Finally the ideas described in the earlier parts of the unit are used in a few of the more common problems in physics.

You will notice the large number of SAQs in this unit. We have set these deliberately, for a considerable amount of the material contained here is revision. We expect you to concentrate on those SAQs which are in the areas where you feel you need most practice, and not to bother with SAQs which test material that you are confident about (although there is no harm in glancing through the solutions even in this case).

6.1 SEPARATION OF VARIABLES

READ W: Section 14, pages 63 to 69.

Notes

(i) *W: page 68, line −17*
The idea of a "closed form" is rather vague, and does not admit a rigorous definition. However, it may be taken to mean an expression involving the elementary functions which can be evaluated using a finite number of arithmetic operations.

(ii) *W: page 68, line −15*
We shall be studying the relevant part of Chapter VII in *Unit 13, Sturm-Liouville Theory*.

(iii) *W: page 69, lines 8 and 9*
This condition amounts to

$$\alpha(y)u(0, y) + [1 - \alpha(y)]\frac{\partial u}{\partial x}(0, y) = 0$$

with $\alpha(y) = 1$ for some values of y and $\alpha(y) = 0$ for the remainder. Thus the coefficients are not independent of y.

General Comment

We have shown that if an equation is separable a solution can be written as a product of functions of one variable. This is obviously of computational convenience, and so it is clearly important to know when an equation will admit a separable solution. It is interesting to note that there is no simple rule for determining whether this is so, and that the best that can be done is to try different variables in turn.

SAQ 1

Assuming that the equation

$$\frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x}\left(x\frac{\partial y}{\partial x}\right)$$

has a solution of the form $y = X(x)T(t)$, determine the equations satisfied by X and T.

(Solution on p. 27.)

SAQ 2

You obtained the two-dimensional Laplace operator in polar coordinates in SAQ 5 of *Unit 3, Elliptic and Parabolic Equations*. Use it to find a solution of

$$r^2\nabla^2\phi = \phi \qquad 0 < r < 1,$$

(where r is the radial coordinate) which

(i) possesses radial symmetry;

(ii) is bounded;

(iii) has the value 1 on the unit circle $r = 1$.

HINT: Look for a solution of the form $\phi = r^\alpha$.

(Solution on p. 27.)

6.2 CONVERGENCE OF FOURIER SERIES

6.2.0 Introduction

Separation of variables leads us to believe that in many cases the solutions to problems involving partial differential equations can be expressed as infinite series. This type of solution first appeared in analysis in connection with the investigations by Daniel Bernoulli (1700–1785) of vibrating cords. He showed in 1748 that a formal solution to the problem of determining the motion of the cord is

$$Y = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l} \cos\frac{n\pi ct}{l},$$

the ends of the cord being fixed at $x = 0$ and l; and he asserted that this was the most general solution to the problem.

A criticism of Bernoulli's theory was published immediately afterwards by Euler, who pointed out that a consequence of the theory was that any function f with $f(0) = f(l) = 0$ could be expressed in the form

$$f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l}.$$

To Euler this was absurd since the series is odd and periodic whilst $f(x)$ need not be. This disagreement led to enquiries into what functions really were and the subsequent investigations touched upon the very foundations of the idea. Eventually this produced the formulation with which you are familiar from the Mathematics Foundation Course.

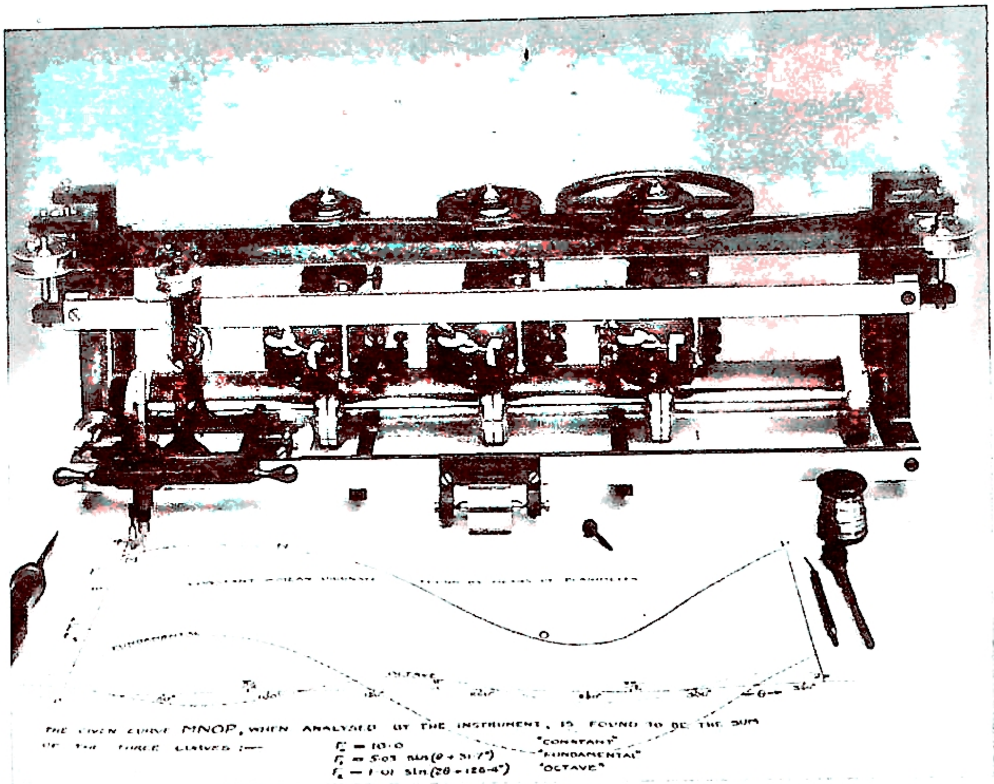
In the eighteenth century, f was called a function only if $f(x)$ could be represented by a simple analytic expression such as a polynomial or a power series. But a mapping with an arbitrary graph, e.g. a polygonal line, was not accepted as a function. The existence of Fourier series, however, showed that such graphs could define functions. It was a long time before this matter was clarified, and the resulting studies were instrumental in the understanding of the foundations of analysis. For example, in 1861 Weierstrass used Fourier series to give an example of a function continuous throughout its domain but without a derivative anywhere.*

Fourier series take their name from Jean Baptiste Joseph Fourier (1768–1830) who based on them his mathematical theory of the conduction of heat presented to the French Academy in 1807. Here he developed the theory of Fourier series and their application to boundary value problems in partial differential equations. He also enounced the proposition that an arbitrary function given graphically by means of a curve, which may be broken by ordinary discontinuities, is capable of representation by means of a trigonometric series. This theorem is said to have been received by Lagrange with astonishment and incredulity.

After this work it was generally agreed that Fourier series were respectable. However, there still remained a few heretics, since Fourier did not provide a proof that the series actually converges at each point to the value of the function concerned. This point was not resolved until almost a century had elapsed.

The importance attached to Fourier series in applied mathematics is evident from the development of a large variety of machines for calculating Fourier components. In the late nineteenth century and the first part of this century there were mechanical analysers working via a complicated system of gears and levers. Some of them can be seen in the Science Museum (London), including Henrici's Analyser of 1894, which is shown in the photograph. Descriptions of many analysers may be found in the literature (see Bibliography). The modern counterparts of these machines are the fast computer algorithms now available.

* This function is quoted in E. C. Titchmarsh, *The Theory of Functions* 2nd ed., (Oxford University Press, 1939), page 351.



Henrici's analyser   Crown copyright, Science Museum.

The idea of representing a function f by an infinite series of known orthogonal functions ϕ_n which form a basis of the function space was introduced in *Unit M201 20, Convergence and Bases*. Such a series is called a *generalized Fourier series*. As was indicated in the Introduction to *Unit M201 22*, the most commonly used basis is the trigonometric basis which produces the (usual) Fourier series. We shall concentrate mainly on this basis, although we shall study first the more general case.

6.2.1 Uniform and Mean Approximation

The discussion of Section 6.1 suggests that in suitable circumstances the solution to a partial differential equation with appropriate boundary conditions may be expressed as an infinite series. The validity of a series solution depends on its convergence. We shall discuss three types of convergence in this section—*pointwise*, *mean* and *uniform*. The discussion of convergence will require some use of techniques from analysis.

We shall meet the three types of convergence in the next reading passage from *W*. However, we first describe uniform convergence in greater detail, since you have probably not met this concept before.

The notion of uniform convergence is important in analysis and has played an important role in the development of the theory of generalized Fourier series. In this course, although it is not a main theme, a rough idea of it will be useful.

SAQ 3

Let $s_n(x)$ be the sum of the finite series

$$x^2 + \frac{x^2}{(1 + x^2)} + \cdots + \frac{x^2}{(1 + x^2)^{n-1}} \qquad x \in R.$$

By noting that this is a geometric series, show that

$$s_n(x) = 1 + x^2 - \frac{1}{(1 + x^2)^{n-1}}.$$

(Solution on p. 27.)

From this SAQ it follows that

$$\lim_{n \to \infty} s_n(x) = 1 + x^2 \qquad x \neq 0.$$

But $s_n(0) = 0$, and so

$$\lim_{n \to \infty} s_n(0) \neq \lim_{x \to 0} \lim_{n \to \infty} s_n(x) = 1,$$

i.e. the two limiting processes as $n \to \infty$ and $x \to 0$ are not interchangeable, despite the fact that $s_n(x)$ is *pointwise* convergent for each $x \in R$.

This strange phenomenon was investigated in the nineteenth century by Stokes, Seidel and Weierstrass who showed that it cannot occur if the sequence of functions converges uniformly; in the above example the sequence $s_1(x), s_2(x), \ldots$ of partial sums does not converge uniformly in the neighbourhood of the origin, as we shall see soon.

The important aspect of the uniform convergence of a sequence $\{f_n\}$ to f (which distinguishes it from mere pointwise convergence) is that given any $\varepsilon > 0$,

$$|f(x) - f_n(x)| < \varepsilon \text{ for all } n > N_\varepsilon \text{ and all } x \text{ in the domain,}$$

where N_ε *is independent of* x. In the example above we showed that for $x \neq 0$ the sequence $\{s_n(x)\}$ converges *pointwise* to $f(x) = 1 + x^2$. By the result of SAQ 3

$$|f(x) - s_n(x)| = \frac{1}{(1 + x^2)^{n-1}},$$

so that given $\varepsilon > 0$,

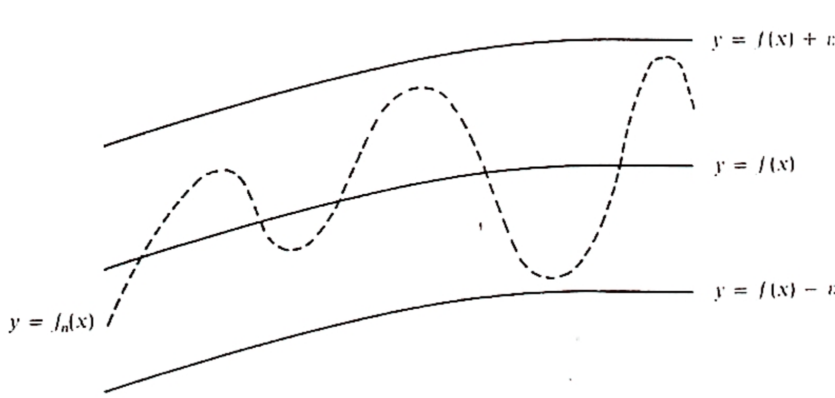
$$|f(x) - s_n(x)| < \varepsilon$$

only if $n > N$, where

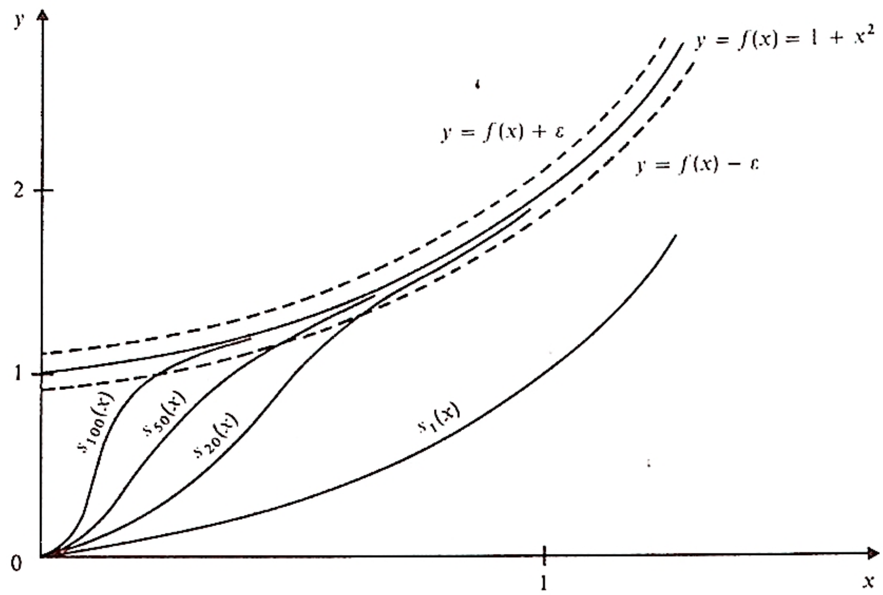
$$N \geqslant 1 - \frac{\ln \varepsilon}{\ln(1 + x^2)}.$$

This expression depends upon x, and is unbounded as $x \to 0$. Thus the convergence of $\{s_n\}$ to f is not uniform. However, the series is uniformly convergent on the domain $\{x \geqslant x_0\}$ where $x_0 > 0$.

We can represent the concept of uniform convergence diagrammatically as follows.



The equivalent situation for the sequence of functions s_n defined in SAQ 3 which does not converge uniformly is shown in the next diagram. It is evident that the sequence is not uniformly convergent on any interval including the origin.



SAQ 4

(i) Show that the sequence

$$f_n(x) = n^2 x(1 - x)^n$$

has the pointwise limit,

$$\lim_{n \to \infty} f_n(x) = 0, \qquad 0 \leqslant x \leqslant 1.$$

(ii) Show that $f_n(x)$ has a local maximum at $x = 1/(1 + n)$ which is unbounded as n becomes large. Hence show that the sequence $\{f_n\}$ does not converge uniformly.

(Solution on p. 28.)

These examples show that uniform convergence is stronger than pointwise convergence, in the sense that the former implies the latter (this follows directly from the definition) but not vice versa.

It is not immediately obvious from the above discussion why uniform convergence is important. There are three main reasons why it is, and we shall make use of them all in this unit.

1 *Continuity*

If a series of continuous functions is uniformly convergent in a given domain, then the sum is also a continuous function.

Thus if u_n $(n = 1, 2, \ldots)$ are continuous functions with domain $[a, b]$, and the series

$$f(z) = \sum_{n=1}^{\infty} u_n(z)$$

is *uniformly* convergent for $a \leqslant z \leqslant b$, then f is continuous on $[a, b]$.

2 *Integration*

If the series

$$f(z) = \sum_{n=1}^{\infty} u_n(z)$$

is *uniformly* convergent for $a \leqslant z \leqslant b$ then the order of integration and summation may be interchanged:

$$\int_a^b f(z)\,dz = \int_a^b \left[\sum_{n=1}^{\infty} u_n(z)\right] dz = \sum_{n=1}^{\infty} \int_a^b u_n(z)\,dz.$$

That is, the series may be integrated term by term as though it were a finite series. (We shall see later that uniform convergence of the series is a sufficient but not necessary condition for this result to hold.)

3 *Differentiation*

If $u_n(n = 1, 2, \ldots)$ are defined on the domain $[a, b]$, and are such that $u_n'(z)$ exist for all z in this domain,

$$\sum_{n=1}^{\infty} u_n(z)$$

converges pointwise in the domain, and the series

$$\sum_{n=1}^{\infty} u_n'(z)$$

is *uniformly* convergent for all z, then the function f given by

$$f(z) = \sum_{n=1}^{\infty} u_n(z) \qquad a \leqslant z \leqslant b,$$

is differentiable and

$$f'(z) = \sum_{n=1}^{\infty} u_n'(z) \qquad a \leqslant z \leqslant b.$$

Thus, if the conditions of the theorem are satisfied the series may be differentiated term by term.

That these results are important when expressing solutions of differential equations in terms of infinite series should not need emphasis. It should be borne in mind, however, that *pointwise* convergence is not sufficient for these results to be valid.

Now we need some criteria indicating when a given series is uniformly convergent. There are two commonly used tests which we describe. We shall be using both of these tests later in this unit.

CAUCHY'S TEST

A necessary and sufficient condition for a sequence of functions u_n with domain I to be uniformly convergent is that given any $\varepsilon > 0$ there exists an integer N_ε independent of $x \in I$ such that

$$|u_n(x) - u_m(x)| < \varepsilon$$

for all $n, m > N_\varepsilon$ and all $x \in I$.

WEIERSTRASS'S M-TEST

This is a sufficient though not necessary condition. If each term of the series

$$f(x) = \sum_{n=1}^{\infty} u_n(x) \qquad x \in I$$

is positive and

$$u_n(x) < M_n \qquad \forall x \in I, n \in Z^+,$$

where each M_n is independent of x, and if

$$\sum_{n=1}^{\infty} M_n$$

is convergent, then the given series for f is uniformly convergent.

We have presented these results in order to use them later. Proofs may be found in Kreider *et al.*, *Linear Analysis*, Appendix I-6. A more detailed discussion is given in Spivak, *Calculus* and *Unit M231 15, Uniform Convergence.*

READ W: Section 15, pages 70 to 72.

Note

W: page 71, line −10
More precisely, eigenfunctions have this property.

General Comment

In general, pointwise convergence does not imply convergence in the mean. This is demonstrated in the next SAQ. Conversely, as we have seen in *Unit M201 20*, mean convergence does not imply pointwise convergence; it follows that mean convergence does not imply uniform convergence. However, it may be shown that a uniformly convergent sequence of functions is convergent in the mean to the same limit.

SAQ 5

(i) Show that the sequence defined by

$$f_n(x) = \begin{cases} 0 & 0 \leqslant x < \dfrac{1}{n+1} \\ \sqrt{n(n+1)} & \dfrac{1}{n+1} \leqslant x \leqslant \dfrac{1}{n} \\ 0 & \dfrac{1}{n} < x \leqslant 1 \end{cases}$$

has the pointwise limit

$$\lim_{n \to \infty} f_n(x) = 0 \qquad 0 \leqslant x \leqslant 1.$$

Thus it is pointwise convergent in [0, 1].

(ii) Show also that

$$\int_0^1 [f_n(x)]^2\,dx = 1$$

for all n, so that the sequence does not converge in the mean to zero with respect to the weight function 1.

(Solution on p. 28.)

SAQ 6 (Optional)

Using the notation of *W: page 72, line 8*, suppose that $f(x) \sim \sum_{n=1}' c_n\phi_n(x)$ where the functions ϕ_n are orthonormal (i.e. orthogonal and such that

$$\int_a^b \phi_n^2 \rho\,dx = 1).$$

Let $s_N(x) = \sum_{n=1}^N c_n\phi_n(x)$, and put $t_N(x) = \sum_{n=1}^N b_n\phi_n(x)$, where the b_n are arbitrary real numbers. Prove that

(a) $$\int_a^b [f(x) - s_N(x)]^2\rho(x)\,dx \leqslant \int_a^b [f(x) - t_N(x)]^2\rho(x)\,dx$$

(this verifies that the c_n, as defined in *W*, provide the best-fitting approximation to f in the function space spanned by $\{\phi_1, \ldots, \phi_N\}$);

(b) if $N > M$,

$$\int_a^b [f(x) - s_N(x)]^2\rho(x)\,dx \leqslant \int_a^b [f(x) - s_M(x)]^2\rho(x)\,dx.$$

(Solution on p. 29.)

SAQ 7

W: page 72, Exercise 1

(Solution on p. 29.)

6.2.2 Parseval's Equation

Having defined convergence, we can now investigate the conditions under which the Fourier series of a function converges. In the next reading passage some properties of the Fourier coefficients are obtained; the main results are conveniently stated as a theorem.

THEOREM

Let $\{\phi_1, \phi_2, \ldots\}$ be orthonormal on $[a, b]$ with respect to the weight function ρ. Suppose that

$$\int_a^b [f(x)]^2 \rho(x)\, dx$$

exists, and that

$$f(x) \sim \sum_{n=1}^{\infty} c_n \phi_n(x).$$

Then the series $\sum_{n=1}^{\infty} c_n^2$ converges and satisfies

$$\sum_{n=1}^{\infty} c_n^2 \leqslant \int_a^b [f(x)]^2 \rho(x)\, dx.$$

This is *Bessel's inequality*. Equality holds if and only if the Fourier series converges to f in the mean. The resulting equation is *Parseval's equation*.

The proof of a more general form of this theorem is contained in the first part of the next reading passage.

READ W: Section 16, pages 73 to 75.

Notes

(i) *W: page 74, line 5*
We note here, without proof, that on the interval $(-\pi, \pi)$ the set of functions $\{1, \sin nx, \cos nx : n = 1, 2, \ldots\}$ forms a complete set. Note also that a complete set is just what was called a *basis* in *Unit M201 20*.

A function f for which $\int_a^b f^2 \rho\, dx$ exists is **square integrable** with weight function ρ.

(ii) *W: page 74, line 6*
The condition that f be continuous has been introduced to eliminate functions of the following type. Suppose that $f(x) = 0$ for all $a \leqslant x \leqslant b$ except at one point where it is non-zero. Then

$$\int_a^b f(x)\phi_n(x)\rho(x)\, dx = 0 \qquad \text{for all } n.$$

Thus this condition does not, on its own, imply that $f \equiv 0$. The requirement that f be continuous allows the implication to follow through.

(iii) *W: page 75, line 6*
A series

$$\sum_{n=1}^{\infty} a_n$$

of real numbers **converges absolutely** if the series

$$\sum_{n=1}^{\infty} |a_n|$$

converges. It follows that a convergent series each of whose terms is positive converges absolutely. It is proved in courses on analysis that if the terms of an absolutely convergent series are rearranged then the resulting series converges

to the same limit as the original series, and that two absolutely convergent series may be subtracted term by term.

SAQ 8

W: page 76, Exercise 1

(Solution on p. 29.)

SAQ 9

W: page 76, Exercise 2

(Solution on p. 30.)

SAQ 10

For a string fixed at $x = 0$ and $x = 1$, and released from rest under the action of no body forces, the most general motion is of the form

$$y = \sum_{n=1}^{\infty} b_n \sin n\pi x \cos n\pi ct,$$

and the nth term of this series is called the nth **harmonic** of y.

Using the expression for the energy in a uniform string of unit length derived in *Unit 2, Classification and Characteristics* (note (i) of Section 2.1.3),

$$E = \tfrac{1}{2}\rho \int_0^1 \left[\left(\frac{\partial y}{\partial t}\right)^2 + c^2 \left(\frac{\partial y}{\partial x}\right)^2 \right] dx,$$

where ρ is the line density of the string, show that Parseval's equation has the interpretation that the total energy of the motion is equal to the sum of the energies in each harmonic.

You may assume that term-by-term differentiation of the series is valid.

(Solution on p. 30.)

6.2.3 Dini's Test

We now proceed to examine conditions for the pointwise convergence of trigonometric Fourier series. The next reading passage shows that the Fourier series of the function f (with respect to the trigonometric basis $\{1, \sin nx, \cos nx : n \in Z^+\}$) converges pointwise to $f(x)$ provided

$$\int_{-\pi}^{\pi} \frac{|f(x+\tau) - f(x)|}{|\tau|} d\tau$$

is finite (i.e. exists). (This is **Dini's test**.) If the one-sided limits of the integrand do not exist as $\tau \rightsquigarrow 0$, then the integral has to be interpreted as

$$\lim_{\varepsilon \to 0^+} \left[\int_{-\pi}^{-\varepsilon} \frac{|f(x+\tau) - f(x)|}{|\tau|} d\tau + \int_{\varepsilon}^{\pi} \frac{|f(x+\tau) - f(x)|}{\tau} d\tau \right].$$

Such an integral is called an **improper integral**.

[The **one-sided limits** $f(x+0)$ and $f(x-0)$ exist, respectively, if $\forall \eta > 0$, $\exists \delta > 0$ such that whenever $\tau \in (0, \delta)$

$$|f(x+\tau) - f(x+0)| < \eta$$

and

$$|f(x-\tau) - f(x-0)| < \eta.$$

In *Unit M201 22* these limits were denoted by $f(x^+)$ and $f(x^-)$. If both one-sided limits exist at x and are equal to $f(x)$ then f is continuous at x; if they are not both equal to $f(x)$ then f has a **jump discontinuity** at x.

A **piecewise continuous** function on $[a, b]$ is continuous at each point of $[a, b]$ except for a finite number of jump discontinuities.]

SAQ 11 (Optional)

Show that if f has a jump discontinuity at x such that $f(x + 0) \neq f(x - 0)$ then

$$\int_{-\pi}^{\pi} \frac{|f(x + \tau) - f(x)|}{|\tau|} d\tau$$

does not exist.

(Solution on p. 31.)

In view of SAQ 11 it is desirable to be able to extend Dini's test to cover piecewise continuous functions, since these frequently occur in practice. In addition, such an extension is required at the end points $\pm\pi$, unless $f(\pi) = f(-\pi)$.

Dini's test is extended (in the next reading passage) to accommodate jump discontinuities in f. At such points the Fourier series converges to

$$\tfrac{1}{2}[f(x + 0) + f(x - 0)]$$

provided the modified Dini's test is satisfied.

We shall also see that for suitable functions f, if $f'(x)$ exists for some $x \in [-\pi, \pi]$ then Dini's test is satisfied at x. Extending this, we could derive the sufficiency condition (which you met in Section 22.2.2 of *Unit M201 22*) that the Fourier series of a *piecewise smooth* function converges pointwise to the average of the left and right limits. This result is the most important to remember; you should not spend too much time on the details in the next reading passage.

READ W_4: Section 18, pages 77 to 80..

Notes

(i) *W: page 79, line 13*
We have not asked you to read *W: Section 17* in which the Riemann–Lebesgue lemma is proved. The Riemann–Lebesgue lemma says that for suitable orthogonal sets of functions $\{\varphi_N\}$

$$\lim_{N\to\infty} \frac{\int_a^b g\phi_N\rho}{\left\{\int_a^b \phi_N^2\rho\right\}^{\frac{1}{2}}} = 0,$$

provided

$$\int_a^b |g|\rho$$

is finite (i.e. the integral exists). In our case

$$a = -\pi,$$

$$b = \pi,$$

$$\rho = 1,$$

$$g(\tau) = \frac{f(x + \tau) - f(x)}{\sin \frac{1}{2}\tau},$$

$$\phi_N(\tau) = \sin(N + \tfrac{1}{2})\tau,$$

and the set $\{\sin(N + \frac{1}{2})\tau : N = 0, 1, 2, \ldots\}$ is orthogonal and "suitable". Since

$$\left\{\int_{-\pi}^{\pi} \sin^2(N + \tfrac{1}{2})\tau \, d\tau\right\}^{\frac{1}{2}} = \pi^{\frac{1}{2}}$$

the condition that

$$\int_{-\pi}^{\pi} \left| \frac{f(x+\tau) - f(x)}{\sin \frac{1}{2}\tau} \right| d\tau \quad \text{, is finite}$$

is sufficient for the right-hand side of Equation (18.6) to approach zero.

(ii) *W: page 79, Equation (18.7)*
The notation

$$< \infty$$

is used to denote that an expression is finite (or that a limit exists).

(iii) *W: page 79, line −13*
The notion of Hölder continuity (more often called the Lipschitz condition of order α) is not important for this course. However, we remark that Hölder continuity implies ordinary continuity, but that the converse is not true (for example, the function $x \longmapsto (1-x)^{-1}$ is continuous on $[0, 1)$ but is not Hölder continuous for any α). We remark too that if $\alpha > 1$, Hölder continuity implies differentiability.

SAQ 12

Why do the Fourier series of the functions $f(x) = x^n$, n being a positive integer, converge pointwise to $f(x)$ for $-\pi < x < \pi$? What happens at $x = \pm\pi$?

(Solution on p. 31.)

SAQ 13

Show that the Fourier series of the function defined on $[-\pi, \pi]$ by

$$\begin{aligned} f(x) &= 0 \qquad -\pi \leqslant x \leqslant 0 \\ &= 1 \qquad 0 < x \leqslant \pi \end{aligned}$$

converges to $f(x)$ everywhere except at $x = 0, \pm\pi$. What is the sum of the series at these points?

(Solution on p. 32.)

In *W: Section 19* it is proved that if f is continuous on $[-\pi, \pi]$, $f(-\pi) = f(\pi)$ and f' is square integrable on $[-\pi, \pi]$ (i.e.

$$\int_{-\pi}^{\pi} f'^2$$

is finite) then the trigonometric Fourier series of f converges *uniformly* to f. This result will be referred to in later reading passages.

Weinberger then goes on to use this property of uniform convergence to prove that the orthogonal set of functions

$$\{1, \sin nx, \cos nx : n = 1, 2, \ldots\}$$

is complete on the interval $[-\pi, \pi]$. This is an important result which you should remember.

We have not asked you to read the proofs in *W: Section 19* since, at this stage, we are rather more concerned in the applications of Fourier series than the theory. Moreover, proofs applicable to a wider range of functions will be given in *Unit 13, Sturm–Liouville Theory*.

A sufficiency condition for a Fourier series to converge *pointwise* to the function it represents is Dini's test. Its main applicability is to functions which are differentiable on $[-\pi, \pi]$.

This concludes our brief analysis of the convergence of Fourier series. We have only skated on the surface of the subject, a complete treatment of which would take at least one full course and a lot of sophisticated analysis.

6.3 GENERALIZED TRIGONOMETRIC SERIES

6.3.0 Introduction

We have seen in the previous section how to form trigonometric Fourier series using the complete orthogonal set

$$\{1, \sin nx, \cos nx : n \in Z^+\}.$$

The importance of such series is evident when we consider the problems in Section 6.1 which led to series expansions with respect to this basis.

We shall now consider several generalizations which extend the use of trigonometric series. The first is the *complex* form which uses the basis $\{e^{inx} : n \in Z\}$ for the vector space of functions $[-\pi, \pi] \longrightarrow C$. This space contains all the real-valued functions on $[-\pi, \pi]$, and indeed the basis functions are closely related to the trigonometric functions by

$$e^{inx} = \cos nx + i \sin nx \qquad n \in Z.$$

We shall also consider sine and cosine series, which make use of the fact that

$$\{\sin nx : n \in Z^+\} \quad \text{and} \quad \{\cos nx : n \in Z_0^+\}$$

are complete orthogonal sets of functions on $[0, \pi]$. Finally we shall investigate trigonometric series on an arbitrary domain $[a, b]$. These last two topics constitute straight revision of material in *Unit M201 22*.

6.3.1 Complex Fourier Series

In this section we shall sketch some results on complex Fourier series, appealing when necessary to the relationship between the complex exponential function and the trigonometric functions given by Euler's formula

$$e^{inx} = \cos nx + i \sin nx$$

(*Unit M100 29, Complex Numbers II*). We obtain from this formula the results

$$\cos x = \frac{1}{2}(e^{ix} + e^{-ix}),$$

$$\sin x = \frac{1}{2i}(e^{ix} - e^{-ix}),$$

$$e^{in\pi} = (-1)^n \qquad n \in Z.$$

Our approach should be reminiscent of *Unit M201 31, Fourier Transforms*, where we used complex-valued functions to unify the handling of the sine and cosine transforms. In order to deal with complex-valued functions it is necessary to define a complex inner product. This is similar to a real inner product. Given a complex vector space V (i.e. a vector space in which multiplication by complex scalars is allowed) and a mapping

$$\cdot : V \times V \longrightarrow C$$

a **complex inner product** is defined on V, if for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$ and $\lambda, \mu \in C$ the following axioms hold.

1 $\quad \mathbf{a} \cdot \mathbf{a} = 0 \Leftrightarrow \mathbf{a} = \mathbf{0}$

2 $\quad \mathbf{a} \cdot \mathbf{a} \geqslant 0$

3 $\quad \mathbf{a} \cdot \mathbf{b} = \overline{\mathbf{b} \cdot \mathbf{a}}$

4 $\quad \mathbf{a} \cdot (\lambda\mathbf{b} + \mu\mathbf{c}) = \lambda\mathbf{a} \cdot \mathbf{b} + \mu\mathbf{a} \cdot \mathbf{c}$

Note how the *conjugate symmetry axiom 3* ensures that $\mathbf{a} \cdot \mathbf{a}$ is real for all $\mathbf{a} \in V$. We are interested in the complex vector space whose elements are functions

$$f : [-\pi, \pi] \longrightarrow C.$$

As usual, we define two vectors **a** and **b** to be **orthogonal** if

$$\mathbf{a} \cdot \mathbf{b} = 0.$$

SAQ 14

Show that the functions

$$x \longmapsto e^{inx} \qquad n \in Z$$

are orthogonal with respect to the inner product

$$\mathbf{f} \cdot \mathbf{g} = \int_{-\pi}^{\pi} \overline{f(x)} g(x)\, dx.$$

(You need not verify that the formula defines a complex inner product.)

(Solution on p. 32.)

The set of orthogonal functions $\{e^{inx}\}$ is equivalent to the trigonometric basis because we have chosen a complex inner product which reduces to our usual real inner product when f and g are real-valued functions, and Euler's formula represents $e^{\pm inx}$ as linear combinations of $\cos nx$ and $\sin nx$, and vice-versa. Consequently, it may be shown that $\{e^{inx} : n \in Z\}$ is complete. We could follow through our discussion of Fourier series for this set of functions, and we would obtain

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}.$$

SAQ 15

Show that

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-inx} f(x)\, dx.$$

(Solution on p. 32.)

The advantages of the complex Fourier series are many. Firstly, it is less space consuming to write down the complex form instead of the real form, and the two calculations for the a_n and b_n are combined in one calculation for the c_n so that only half of the work is required; we shall see an example of this in the solution to SAQ 18. Secondly, under the operations of differentiation and integration e^{inx} is easier to handle than either $\sin nx$ or $\cos nx$. Thus,

$$\begin{aligned}
\frac{d^p}{dx^p}(\sin nx) &= (-1)^q n^p \sin nx \quad \text{if} \quad p = 2q \\
&= (-1)^q n^p \cos nx \quad \text{if} \quad p = 2q + 1 \\
\frac{d^p}{dx^p}(\cos nx) &= (-1)^q n^p \cos nx \quad \text{if} \quad p = 2q \\
&= (-1)^q n^p \sin nx \quad \text{if} \quad p = 2q - 1 \\
\frac{d^p}{dx^p}(e^{inx}) &= (in)^p e^{inx}
\end{aligned}$$

with similar results for integration. The usefulness of this simplification where dealing with differential equations need hardly be mentioned.

There are other reasons why the complex form of Fourier series is used. Most of the theory of real analysis, of which Fourier series is a part, is easier to understand when viewed from the more general complex variable theory and in this theory the exponential function is a fundamental function.

In obtaining these simplifications we have introduced the inconvenience of complex numbers; but when one is used to their manipulations they avoid the pitfalls encountered when handling sines and cosines, which are prone, at the least, to getting their signs muddled!

SAQ 16

If

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}$$

and

$$f(x) \sim \frac{a_0}{2} + \sum_1^{\infty} (a_n \cos nx + b_n \sin nx)$$

show that

$$\left.\begin{aligned} c_n &= \tfrac{1}{2}(a_n - ib_n) \\ c_{-n} &= \tfrac{1}{2}(a_n + ib_n) \end{aligned}\right\} \quad n > 0,$$
$$c_0 = \tfrac{1}{2}a_0.$$

(Solution on p. 32.)

SAQ 17

Show that if f is a real-valued function then

$$\bar{c}_n = c_{-n}.$$

(Solution on p. 33.)

6.3.2 Sine and Cosine Series

The gist of this section is that an even (odd) function can be expressed as a sum of even (odd) functions. Since cos is an even function and sin is an odd function the Fourier series for an even (odd) function contains only cosines (sines).

READ W: Section 20, pages 87 to 88.

Note

W: page 87, lines −6 and −5
The reference to Parseval's equation is unnecessary. Since $\{1, \sin nx, \cos nx : n = 1, 2, \ldots\}$ is complete (as proved in *W: Section 19*) any suitably behaved function can be expressed in terms of these functions with Fourier components given by Equations (18.1) in *W: Page 78*. If the function is odd, then $a_n = 0$ for all n, and it follows that any suitably behaved odd function may be expressed in terms of $\sin nx$ $(n = 1, 2, \ldots)$ only. Thus $\{\sin nx : n = 1, 2, \ldots\}$ is complete on the space of odd functions with domain $[-\pi, \pi]$. Since any function with domain $[0, \pi]$ can be extended as an odd function on $[-\pi, \pi]$, $\{\sin nx\}$ is complete on the space of (square integrable) functions with domain $[0, \pi]$.

SAQ 18

W: page 88, Exercise 1

(Solution on p. 33.)

6.3.3 Change of Scale

READ W: Section 21, pages 88 to 92.

SAQ 19

W: page 92, Exercise 3

(Solution on p. 34.)

SAQ 20

W: page 92, Exercise 5

(Solution on p. 35.)

6.4 APPLICATIONS

6.4.0 Introduction

We now proceed to apply the methods discussed in the previous sections to problems involving partial differential equations. Many of these applications were introduced in *Unit M201 32*, where both the heat equation and Laplace's equation were solved for a variety of boundary conditions. The treatment presented here is slightly different, much more effort being devoted to rigour.

The problems which concern us involve a partial differential equation in some domain D with a continuous boundary condition on its boundary C. In the case of the heat equation, C may be taken as the physical boundary for $t > 0$ together with the whole initial line. A solution of the problem must satisfy the equation and "boundary" conditions and be continuous on $D \cup C$.

In *W: Sections 22 and 23* the following method is used. Firstly, the method of separation of variables yields a Fourier series solution which *formally* satisfies the differential equation and subsidiary conditions: it is still necessary to prove that this series does indeed converge to a solution. Next, Weierstrass's M-test (Section 6.2.1) is used to show that the series is uniformly convergent on part of $D \cup C$. This ensures that the sum u of the series is continuous in the relevant part of $D \cup C$. Next, it is shown that the series obtained by differentiating term by term at any point in D is uniformly convergent in some subset of D, so that u really satisfies the differential equation. Finally, Cauchy's test (Section 6.2.1) and the maximum principle are used to ensure uniform convergence throughout $D \cup C$ (provided some extra conditions are satisfied) so that u is continuous everywhere.

Note that in *Unit 3* we have already proved, using the maximum principle, that suitable subsidiary conditions for the heat equation and Laplace's equation yield *uniqueness* of the solution and its *continuity with respect to the data*. This section proves the existence of solutions to selected problems of this type and thus shows that such problems are *properly posed*. This is as we would expect, since the problems are based on realistic physical situations.

6.4.1 The Heat Equation

Our first example is the one-dimensional heat equation with homogeneous boundary conditions.

READ W: Section 22, pages 92 to 95.

Notes

(i) *W: page 93, lines 15 to 17*

The statement that $\{b_n\}$ is **uniformly bounded** means that there exists $c \in R$, independent of n, such that, for all n, $|b_n| < c$.

The uniform bound is obtained by using the result

$$\left|\int_a^b g(x)\,dx\right| \leqslant \int_a^b |g(x)|\,dx$$

which is the integral equivalent of the triangle rule for finite sums,

$$|a_1 + a_2 + \cdots + a_n| \leqslant |a_1| + |a_2| + \cdots + |a_n|.$$

Then since $|\sin nx| \leqslant 1$, $|f(x) \sin nx| \leqslant |f(x)|$ and

$$|b_n| = \left|\frac{2}{\pi}\int_0^\pi f(x)\sin nx\,dx\right| \leqslant \frac{2}{\pi}\int_0^\pi |f(x)|\,dx.$$

(ii) *W: page 93, lines −11 to −4*
You need not verify that these series converge for $t > 0$. However, this said, it is obvious that the convergence is uniform in x. (To show that it is uniform for $t \geqslant t_0$, we require the M-test with $M_n = \exp(-n^2kt_0)$ and similarly for the other series.)

(iii) *W: page 94, lines 10 to 12*
This follows because the Fourier sine series of f is uniformly convergent.

SAQ 21

Find the analytical solution of

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad 0 < x < 1, t > 0.$$

$$u(0, t) = u(1, t) = 0 \qquad t > 0,$$

$$u(x, 0) = \begin{cases} 2x & 0 \leqslant x \leqslant \frac{1}{2} \\ 2(1 - x) & \frac{1}{2} \leqslant x \leqslant 1. \end{cases}$$

We discussed the numerical solution to this problem in *Unit 5, Initial Value Problems* (*S: pages 11 to 16*).

(Solution on p. 35.)

SAQ 22

Write down the heat equation and boundary conditions for a thin conducting ring or armlet of circumference $2l$ if the initial temperature is $\sin^2(\pi x/2l)$, where x is the distance around the circumference from some fixed point. Determine the subsequent temperature distribution.

This problem is especially interesting as it was one of the first to which Fourier applied his mathematical theory of heat, and for which the results of his mathematical investigations were compared with experiment. For the original theory see Fourier, *The Analytical Theory of Heat*.

(Solution on p. 36.)

6.4.2 Laplace's Equation

READ W: Section 23, pages 95 to 98.

If you are short of time you may omit the discussion of error bounds, *W: page 97, line −11 to page 98, line 7.*

Note

W: page 98, final paragraph
Section 25 is not included in this course.

As an illustration of this method of solving Laplace's equation we consider the problem discussed in *Unit 3*, that is, the calculation of the total rate of steady flow, Q, of a viscous fluid through a pipe of square cross section. In Section 3.2.3 we showed that

$$0{\cdot}555 \leqslant \frac{Q}{ka^4} \leqslant 0.667,$$

where $2a$ is the length of a side of the square and k a constant (defined in Section 3.2.1). Here we determine the exact value of Q which is given by

$$Q = \int_{-a}^{a} \int_{-a}^{a} w \, dx \, dy$$

where w is the velocity of the fluid and satisfies the equation

$$\nabla^2 w = -k,$$

with $w = 0$ on the boundary. This equation is not quite of the form we require, but on defining

$$\phi(x, y) = w + \frac{k}{4}(x^2 + y^2 - 2a^2)$$

we find that

$$\nabla^2 \phi = 0$$

and that

$$\phi(x, y) = \frac{k}{4}(x^2 + y^2 - 2a^2) \qquad \text{on the boundary.}$$

The choice of ϕ satisfying $\nabla^2 \phi = 0$ is not unique; any function of the form

$$(x, y) \longmapsto w + \frac{k}{4}(x^2 + y^2) + \alpha x + \beta y + \gamma xy + \delta$$

where α, β, γ and δ are arbitrary constants will do (since the second term goes to k under the operator ∇^2, and the remaining terms go to zero). The reason for our choice will soon be apparent.

Although now the equation to be solved is simpler the boundary conditions are more complicated. In order to solve this problem we use the technique of *W: page 98* (also discussed in *W: page 64*) and put ϕ equal to $\phi_1 + \phi_2 + \phi_3 + \phi_4$, each of which satisfies $\nabla^2 \phi_i = 0$ and is nonzero on only one side of the square. This apparently necessitates solving four equations. But the symmetry of the problem shows that only ϕ_1 need be found; all the others can be obtained by either changing signs, or interchanging x and y. Thus we need to solve

$$\nabla^2 \phi_1 = 0 \qquad x, y \in (-a, a),$$

$$\phi_1(x, a) = \frac{k}{4}(x^2 - a^2) \qquad x \in [-a, a],$$

$$\phi_1(x, -a) = 0 \qquad x \in [-a, a],$$

$$\phi_1(a, y) = \phi_1(-a, y) = 0 \qquad y \in [-a, a].$$

Now it is clear how the arbitrary constants of ϕ were chosen. With our choice of constants, $\phi(x, y) = 0$ at the corners of the square so that each ϕ_i is continuous on the boundary. With any other choice of constants ϕ_i would not be continuous on the boundary, and discontinuities are best avoided as the Fourier series of discontinuous functions are not uniformly convergent over the whole interval, and also converge more slowly, creating a new computational inconvenience.

The solution to this problem is now straightforward but lengthy. By separating the variables and using the boundary conditions along $x = \pm a$ and $y = -a$ we find that

$$\phi_1(x, y) = \sum_{n=1}^{\infty} c_n \sinh \frac{n\pi(y + a)}{2a} \sin \frac{n\pi(x + a)}{2a}.$$

At $y = a$,

$$\phi_1(x, a) = \frac{k}{4}(x^2 - a^2) = \sum_{n=1}^{\infty} c_n \sinh n\pi \sin \frac{n\pi(x + a)}{2a} \qquad x \in [-a, a],$$

so that

$$c_n \sinh n\pi = \frac{1}{a}\int_{-a}^{a} \frac{k}{4}(x^2 - a^2) \sin \frac{n\pi(x + a)}{2a}\, dx$$

$$= \frac{4a^2 k}{n^3 \pi^3}((-1)^n - 1).$$

Since the boundary condition is continuous, the series for ϕ_1 is uniformly convergent in $[-a, a] \times [-a, a]$ and the differentiated series are uniformly convergent in $(-a, a) \times (-a, a)$, as in the reading passage.

The total flux is now given by

$$Q = \int_{-a}^{a}\int_{-a}^{a}\left[\phi_1(x, y) + \phi_2(x, y) + \phi_3(x, y) + \phi_4(x, y) - \frac{k}{4}(x^2 + y^2 - 2a^2)\right] dx\, dy$$

$$= 4\int_{-a}^{a}\int_{-a}^{a} \phi_1(x, y)\, dx\, dy + \frac{4ka^4}{3}$$

$$= \frac{4ka^4}{3} + 4\sum_{n=1}^{\infty} c_n \int_{-a}^{a} \sin\frac{n\pi(x + a)}{2a}\, dx \int_{-a}^{a} \sinh\frac{n\pi(y + a)}{2a}\, dy$$

$$= ka^4\left[\frac{4}{3} - \frac{256}{\pi^5}\sum_{m=0}^{\infty}\frac{\cosh(2m + 1)\pi - 1}{(2m + 1)^5 \sinh(2m + 1)\pi}\right].$$

Using the first three terms, this gives

$$Q = 0.562ka^4$$

to three places of decimals.

SAQ 23

W: page 99, Exercise 9

(Solution on p. 37.)

Laplace's equation may be solved for a circular boundary by methods similar to those of this section. The discussion appears in *W: Section 24* which we shall cover in *Unit 10, Green's Functions II*.

6.5 SUMMARY

In this unit we have considered two main topics, both of which were first introduced in M201.

We showed how some partial differential equations could be solved by *separation of variables*. We saw that equations which permitted this treatment led to *Fourier series*; and we found that certain boundary value problems and initial-boundary value problems yielded solutions under this treatment.

In considering the theory of Fourier series we introduced the notions of *uniform convergence*, *convergence in the mean* and *pointwise convergence*: the ideas of *orthogonal*, *orthonormal* and *complete* sets of functions were also used. A new result, *Dini's Test*, was obtained as a sufficient condition for the pointwise convergence of a trigonometric series. We also obtained a general form of *Parseval's equation* which implies that a Fourier series (with respect to any complete orthogonal set of functions) can be integrated term by term to yield a pointwise convergent series.

These techniques were applied to a few physical examples, using some general theorems on uniform convergence and a result (which we did not prove) concerning sufficient conditions for a trigonometric Fourier series to converge uniformly.

A brief discussion illustrated the use of complex exponential functions as a complete set of orthogonal functions which yield the complex generalization of trigonometric Fourier series.

6.6 FURTHER SELF-ASSESSMENT QUESTIONS

SAQ 24

Show that, for $\alpha > -\frac{1}{2}$,

$$\lim_{n \to \infty} \int_0^{\pi} x^{\alpha} \sin nx \, dx = 0.$$

(Solution on p. 38.)

SAQ 25

The function f is odd and for $0 < x < \pi$, $f(x) = \cos x$. Obtain the Fourier series for f in the form

$$f(x) \sim \frac{8}{\pi} \sum_{m=1}^{\infty} \frac{m \sin 2mx}{4m^2 - 1}.$$

(Solution on p. 38.)

SAQ 26

A function f with period 2π is equal to x^2 for $-\pi \leqslant x \leqslant \pi$. Show that

$$f(x) = \frac{\pi^2}{3} + 4 \sum_{n=1}^{\infty} \frac{\cos n\pi \cos nx}{n^2} \qquad x \in R.$$

Deduce that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

(Solution on p. 39.)

SAQ 27

Is the function

$$f: x \longmapsto |x|^{-\alpha} \qquad x \in (-\pi, 0) \cup (0, \pi)$$

piecewise smooth for $\alpha > 0$? Show that if $0 < \alpha < 1$ then the Fourier series for f converges pointwise to f for $x \neq 0$.

(Solution on p. 39.)

6.7 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

Substituting $y = X(x)T(t)$ into the equation we obtain

$$X(x)\frac{d^2T}{dt^2}(t) = T(t)\frac{d}{dx}\left(x\frac{dX}{dx}(x)\right)$$

which becomes, on rearrangement,

$$\frac{1}{T(t)}\frac{d^2T}{dt^2}(t) = \frac{1}{X(x)}\frac{d}{dx}\left(x\frac{dX}{dx}(x)\right).$$

Since the left-hand side depends on t only, and the right-hand side is independent of t, each side must be equal to a constant, α say. Thus the required equations are

$$\frac{d^2T}{dt^2} - \alpha T = 0,$$

$$x\frac{d^2X}{dx^2} + \frac{dX}{dx} - \alpha X = 0.$$

Solution to SAQ 2

Since the solution possesses radial symmetry the most suitable coordinates are polar coordinates (r, θ) in which the solution must be independent of the polar angle θ. The equation is thus

$$r^2\frac{\partial^2\phi}{\partial r^2} + r\frac{\partial\phi}{\partial r} + \frac{\partial^2\phi}{\partial\theta^2} = \phi.$$

Since ϕ has axial symmetry, $\partial\phi/\partial\theta = 0$ and we may write

$$r^2\frac{d^2\phi}{dr^2} + r\frac{d\phi}{dr} = \phi \qquad 0 < r < 1, \tag{*}$$

where ϕ depends on r only.

We look for a solution of the form $\phi = r^\alpha$ where α is some constant; on substituting into the equation we find that

$$(\alpha^2 - 1)r^\alpha = 0.$$

Since $r^\alpha \neq 0$, $\alpha = \pm 1$, and the solution space contains r^{-1} and r. Since the equation (*) is normal† and second-order, its solution space is two-dimensional. So the general solution is

$$\phi = A/r + Br.$$

But the solution is bounded in $(0, 1)$, and this can only be so if $A = 0$: also, at $r = 1$ we have $\phi = B = 1$, giving

$$\phi = r$$

as the required solution.

Solution to SAQ 3

The sum of a geometric series is given by

$$a\sum_{p=0}^{n-1} r^p = a\frac{1 - r^n}{1 - r}.$$

In this case $a = x^2$, $r = 1/(1 + x^2)$; thus

$$s_n(x) = 1 + x^2 - \frac{1}{(1 + x^2)^{n-1}}.$$

† A differential equation is normal on the domain I if its leading coefficient is nonzero throughout I (*Unit M201 9, Differential Equations II*).

Solution to SAQ 4

(i) For $x = 0, 1$ we have $f_n(0) = f_n(1) = 0$ for each $n \in Z^+$, so that

$$\lim_{n \to \infty} f_n(x) = 0 \qquad x = 0, 1.$$

Let $\xi = 1 - x$. If $0 < x < 1$ then $0 < \xi < 1$, so that

$$0 < f_n(x) = xn^2\xi^n < n^2\xi^n \qquad 0 < x < 1$$

and, since $|\xi| < 1$,

$$\lim_{n \to \infty} n^2\xi^n = 0.$$

(To verify this statement, note that $0 < \xi^{\frac{1}{3}} < 1$ so that we may choose an integer N such that

$$\xi^{\frac{1}{3}} < \frac{N}{N+1}.$$

For $n \geqslant N$ the sequence $n^3\xi^n$ is decreasing, i.e.

$$(n+1)^3\xi^{n+1} < n^3\xi^n < N^3\xi^N = k, \text{ say.}$$

Thus

$$0 \leqslant \lim_{n \to \infty} n^2\xi^n \leqslant \lim_{n \to \infty} \frac{k}{n} = 0.)$$

Hence

$$\lim_{n \to \infty} f_n(x) = 0 \qquad 0 \leqslant x \leqslant 1.$$

(ii) The stationary points of $f_n(x)$ are given by the roots of $f_n'(x) = 0$, i.e.

$$f_n'(x) = n^2\big((1 - x) - xn\big)(1 - x)^{n-1} = 0.$$

Thus $f(x)$ has a stationary point at $x = 1/(1 + n)$ which can be seen to be a maximum by inspection. At this point

$$f_n(x) = \frac{n^2}{1+n}\frac{1}{(1 + (1/n))^n}$$

Since

$$\lim_{n \to \infty}\left(1 + \frac{1}{n}\right)^n = e$$

(*Unit M100 7, Sequences and Limits I*), $f_n(1/(1 + n))$ behaves as n and so is unbounded.

Solution to SAQ 5

(i) For pointwise convergence to 0 we require, for each x in $[0, 1]$, that:

given any $\varepsilon > 0, \exists N$ such that $|f_n(x)| < \varepsilon$ for all $n > N$.

Now, given any x in $(0, 1]$ we can always find an $N > 1$ such that $x > 1/N$ (take N to be the first integer $> 1/x$). Then, for $n > N, f_n(x) = 0$ for $0 < x \leqslant 1$. It follows that

$$\lim_{n \to \infty} f_n(x) = 0 \qquad 0 < x \leqslant 1,$$

and from the definition $f_n(0) = 0 \; \forall n$, so that the result follows.

(ii) We have

$$\int_0^1 [f_n(x)]^2 \, dx = \int_{1/n+1}^{1/n} n(n+1)\, dx$$

$$= n(n+1)\left(\frac{1}{n} - \frac{1}{n+1}\right) = 1.$$

This SAQ illustrates the point that pointwise convergence does not imply convergence in the mean.

Solution to SAQ 6

(a) Expanding the right-hand side and completing the square as in *W: page 71*, and remembering that $\int_a^b \phi_n^2 \rho \, dx$ is unity since the ϕ_n are orthonormal, we obtain

$$\int_a^b [f(x) - t_N(x)]^2 \rho(x) \, dx$$

$$= \sum_{n=1}^{N} \left\{ b_n - \int_a^b f\phi_n \rho \, dx \right\}^2 + \int_a^b f^2 \rho \, dx - \sum_{n=1}^{N} \left\{ \int_a^b f\phi_n \rho \, dx \right\}^2$$

$$= \sum_{n=1}^{N} (b_n - c_n)^2 + \int_a^b f^2 \rho \, dx - \sum_{n=1}^{N} c_n^2 \text{ by the definition of } c_n.$$

Thus,

$$\int_a^b [f(x) - s_N(x)]^2 \rho(x) \, dx = \int_a^b f^2 \rho \, dx - \sum_{n=1}^{N} c_n^2,$$

so that

$$\int_a^b [f(x) - s_N(x)]^2 \rho(x) \, dx = \int_a^b [f(x) - t_N(x)]^2 \rho(x) \, dx - \sum_{n=1}^{N} (b_n - c_n)^2,$$

and the result follows.

(b) This result follows directly from part (a) by putting

$$b_i = c_i \qquad i = 1, 2, \ldots M,$$
$$b_i = 0 \qquad i = M + 1, \ldots N,$$

so that $t_N(x) = s_M(x)$.

Solution to SAQ 7

We need to show that

$$\int_0^\pi \sin nx \sin mx \, dx = 0 \qquad n \neq m, \quad n, m \in Z^+.$$

The integral may be written in the form

$$\frac{1}{2}\int_0^\pi (\cos(n - m)x - \cos(n + m)x) \, dx$$

$$= \frac{1}{2}\left[\frac{\sin(n - m)x}{n - m} - \frac{\sin(n + m)x}{n + m}\right]_0^\pi \qquad n \neq m$$

$$= 0.$$

For $n = m$ the integral is

$$\int_0^\pi \sin^2 nx \, dx = \frac{\pi}{2}.$$

Solution to SAQ 8

The Fourier sine series for

$$f(x) = 1 \qquad 0 \leqslant x \leqslant \pi$$

is

$$f(x) \sim \sum_{n=1}^{\infty} d_n \sin nx,$$

where the Fourier coefficients are given by

$$d_n = \frac{\int_0^\pi \sin nx\,dx}{\int_0^\pi \sin^2 nx\,dx} = \frac{\frac{1}{n}\Big[-\cos nx\Big]_0^\pi}{\pi/2}$$

$$= \begin{cases} 0 & n \text{ even} \\ \dfrac{4}{\pi n} & n \text{ odd.} \end{cases}$$

Thus

$$1 \sim \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{\sin(2k-1)x}{2k-1} \qquad x \in [0, \pi].$$

If the set of functions $\{\sin nx\}$ is complete on $[0, \pi]$ we may use the result of Equation (16.8) in *W: page 75* to integrate term by term, obtaining

$$\int_0^x dx' = \frac{4}{\pi}\sum_{k=1}^{\infty}\int_0^x \frac{\sin(2k-1)x'}{2k-1}\,dx'$$

or

$$x = \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{1-\cos(2k-1)x}{(2k-1)^2} \qquad x \in [0, \pi].$$

The equality here denotes pointwise convergence.

Solution to SAQ 9

Parseval's equation gives

$$\int_0^\pi dx = \frac{16}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\int_0^\pi \sin^2(2k-1)x\,dx,$$

i.e.

$$\pi = \frac{8}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}.$$

Substitution of π into the series for x in the solution to SAQ 8 yields the same result:

$$\pi = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{2}{(2k-1)^2},$$

i.e.

$$\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2} = \frac{\pi^2}{8}.$$

Solution to SAQ 10

The energy of a uniform string of unit length is given as

$$E = \tfrac{1}{2}\rho\int_0^1 \left[\left(\frac{\partial y}{\partial t}\right)^2 + c^2\left(\frac{\partial y}{\partial x}\right)^2\right] dx.$$

The motion is of the form

$$y = \sum_{n=1}^{\infty} b_n \sin n\pi x \cos n\pi c t \qquad 0 \leqslant x \leqslant 1,\ t \geqslant 0,$$

and the nth harmonic is

$$y_n = b_n \sin n\pi x \cos n\pi c t.$$

The energy E_n in this harmonic is obtained simply by substituting y_n in the above expression for the energy:

$$E_n = \tfrac{1}{2}\rho b_n^2(n\pi c)^2 \int_0^1 (\sin^2 n\pi x \sin^2 n\pi ct + \cos^2 n\pi x \cos^2 n\pi ct)\,dx$$

$$= \tfrac{1}{4}\rho(n\pi c b_n)^2 \quad \text{since} \quad \int_0^1 \sin^2 n\pi x\,dx = \int_0^1 \cos^2 n\pi x\,dx = \tfrac{1}{2}.$$

Now, term-by-term differentiation yields the Fourier series

$$\frac{\partial y}{\partial t} = -\sum_{n=1}^{\infty} n\pi c b_n \sin n\pi x \sin n\pi ct$$

and

$$\frac{\partial y}{\partial x} = \sum_{n=1}^{\infty} n\pi b_n \cos n\pi x \cos n\pi ct,$$

and applying Parseval's equation, we obtain

$$\int_0^1 \left(\frac{\partial y}{\partial t}\right)^2 dx = \sum_{n=1}^{\infty} (n\pi c b_n)^2 \sin^2 n\pi ct \int_0^1 \sin^2 n\pi x\,dx$$

and

$$\int_0^1 \left(\frac{\partial y}{\partial x}\right)^2 dx = \sum_{n=1}^{\infty} (n\pi b_n)^2 \cos^2 n\pi ct \int_0^1 \cos^2 n\pi x\,dx.$$

The total energy is then

$$E = \tfrac{1}{4}\rho \sum_{n=1}^{\infty} (n\pi c b_n)^2 \quad \text{since} \quad \int_0^1 \sin^2 n\pi x\,dx = \int_0^1 \cos^2 n\pi x\,dx = \tfrac{1}{2}$$

$$= \sum_{n=1}^{\infty} E_n.$$

Thus the total energy of a vibrating string (or more generally a vibrating system) is the sum of the energies in each harmonic.

Solution to SAQ 11

Let $\eta = \frac{1}{5}|f(x+0) - f(x-0)|$; suppose that

$$|f(x+0) - f(x)| > 2\eta.$$

(If not, then

$$|f(x-0) - f(x)| > 2\eta,$$

and a similar argument will follow.) Now $\exists\delta > 0$ such that

$$\forall \tau \in (0, \delta) \qquad |f(x+\tau) - f(x+0)| < \eta.$$

Hence,

$$\forall \tau \in (0, \delta) \qquad |f(x+\tau) - f(x)| > \eta > 0.$$

Therefore

$$\lim_{\varepsilon \to 0^+} \int_\varepsilon^\delta \frac{|f(x+\tau) - f(x)|}{|\tau|}\,d\tau > \eta \lim_{\varepsilon \to 0^+} \int_\varepsilon^\delta \frac{1}{\tau}\,d\tau = \eta \lim_{\varepsilon \to 0^+} \ln(\delta/\varepsilon)$$

which does not exist. Hence,

$$\int_{-\pi}^{\pi} \frac{|f(x+\tau) - f(x)|}{|\tau|}\,d\tau$$

does not exist.

Solution to SAQ 12

Because the functions $f(x) = x^n$ are differentiable for $-\pi < x < \pi$ and absolutely integrable.

For n even $f(\pi) = f(-\pi)$, with the consequence that the Fourier series converges pointwise to $f(x)$ for $-\pi \leqslant x \leqslant \pi$. If n is odd $f(\pi) \neq f(-\pi)$, and the Fourier series converges to zero for $x = \pm\pi$, since $\frac{1}{2}[f(\pi) + f(-\pi)] = 0$.

Solution to SAQ 13

The function is piecewise smooth and so at each interior point converges to

$$\frac{f(x+0)+f(x-0)}{2}=\begin{cases}0 & -\pi<x<0\\ 1 & 0<x<\pi\\ \frac{1}{2} & x=0,\end{cases}$$

whilst at $x = \pm\pi$ it converges to

$$\frac{f(-\pi+0)+f(\pi-0)}{2}=\frac{1}{2}.$$

Solution to SAQ 14

The solution to this SAQ is similar to that of SAQ 7. We need to evaluate the integral

$$I=\int_{-\pi}^{\pi} e^{ix(n-m)}\,dx \qquad n, m \in Z.$$

For $n = m$ clearly $I = 2\pi$, since $e^{i0} = 1$. For $n \neq m$

$$\begin{aligned} I &= \frac{1}{i(n-m)}\Big[e^{ix(n-m)}\Big]_{-\pi}^{\pi} \\ &= \frac{1}{i(n-m)}\left(e^{i\pi(n-m)} - e^{-i\pi(n-m)}\right) \\ &= \frac{e^{i\pi(n-m)}}{i(n-m)}\left(1 - e^{-2i\pi(n-m)}\right) = 0, \end{aligned}$$

since $e^{2i\pi p} = 1$ when p is an integer.

Solution to SAQ 15

The main difference between our case and the discussion in *W: Section 15* is that the exponential functions are orthogonal with respect to the complex inner product defined in SAQ 14. (The complex inner product is more general than the real one defined in *W*.) Following through the analysis of *W: Section 15* with our complex inner product then gives

$$c_n=\frac{\displaystyle\int_{-\pi}^{\pi} f(x)\overline{\phi_n(x)}\,dx}{\displaystyle\int_{-\pi}^{\pi}|\phi_n(x)|^2\,dx}.$$

In our case $\phi_n(x) = e^{inx}$, so that

$$c_n=\frac{\displaystyle\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx}{\displaystyle\int_{-\pi}^{\pi}dx}=\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx.$$

Solution to SAQ 16

We have

$$c_n=\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx \qquad n\in Z \qquad \text{(SAQ 15)}$$

and

$$\left.\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \qquad n\in Z_0^+ \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \qquad n\in Z^+ \end{aligned}\right\} \quad \text{(Equation (18.1) in } W\text{: page 78).}$$

Thus, for $n > 0$

$$\tfrac{1}{2}(a_n - ib_n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos nx - i\sin nx)\,dx$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx$$

$$= c_n$$

and

$$\tfrac{1}{2}(a_n + ib_n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos nx + i\sin nx)\,dx$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{inx}\,dx$$

$$= c_{-n}.$$

Also

$$\tfrac{1}{2}a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx$$

$$= c_0.$$

Solution to SAQ 17

Since

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx,$$

we have

$$\overline{c_n} = \frac{1}{2\pi}\overline{\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \overline{f(x)e^{-inx}}\,dx$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{inx}\,dx \quad \text{since } f(x) \text{ is real}$$

$$= c_{-n}.$$

Solution to SAQ 18

Both of the required series may be found in one calculation if the function

$$f(x) = x \qquad 0 \leqslant x \leqslant \pi$$

is extended into

$$[-\pi, \pi]$$

in the correct manner. For the sine series we note that $\sin nx$ is odd and so we extend f so that it is odd, i.e.

$$f_o(x) = x \qquad -\pi \leqslant x \leqslant \pi.$$

The cosine series is even and so we extend f so that it is even, i.e.

$$f_e(x) = |x| \qquad -\pi \leqslant x \leqslant \pi.$$

If we now find the Fourier series of $f_o(x)$ and $f_e(x)$ on $-\pi \leqslant x \leqslant \pi$, they will be the required sine and cosine series for

$$f(x) = x \qquad 0 \leqslant x \leqslant \pi.$$

Thus we have

$$x \sim \sum_{n=1}^{\infty} b_n \sin nx$$

where

$$b_n = \frac{2}{\pi}\int_0^{\pi} x \sin nx \, dx,$$

and

$$|x| \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx$$

where

$$a_n = \frac{2}{\pi}\int_0^{\pi} x \cos nx \, dx.$$

Writing

$$c_n = a_n + ib_n = \frac{2}{\pi}\int_0^{\pi} xe^{inx} \, dx,$$

and integrating by parts, we find that

$$c_n = \frac{2}{\pi}\left[\frac{\pi(-1)^n}{in} + \frac{1}{n^2}((-1)^n - 1)\right] \qquad n > 0,$$

and so

$$a_n = \operatorname{Re} c_n = \frac{2}{\pi}\frac{(-1)^n - 1}{n^2}, \qquad a_0 = \pi,$$

$$b_n = \operatorname{Im} c_n = \frac{-2(-1)^n}{n}.$$

Thus

$$x \sim -2\sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin nx \quad \text{on} \quad [0, \pi]$$

and

$$x \sim \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=1}^{\infty} \frac{\cos(2k-1)x}{(2k-1)^2} \quad \text{on} \quad [0, \pi].$$

Solution to SAQ 19

Let x and t be the space and time coordinates for the first bar. The heat equation describing its temperature $u(x, t)$ is

$$\frac{\partial u}{\partial t} - k\frac{\partial^2 u}{\partial x^2} = 0 \qquad 0 < x < l, \quad t > 0.$$

$f(t) = u(\frac{1}{2}l, t)$ is the temperature at the centre of this bar at time t. Putting

$$\bar{x} = x\frac{\bar{l}}{l}, \; \bar{t} = t\frac{\bar{l}^2}{l^2}$$

the heat equation becomes

$$\frac{\partial \bar{u}}{\partial \bar{t}} - k\frac{\partial^2 \bar{u}}{\partial \bar{x}^2} = 0 \qquad 0 < \bar{x} < \bar{l}, \quad \bar{t} > 0,$$

where $\bar{u}(\bar{x}, \bar{t}) = u(x, t)$. This is just the equation satisfied by the temperature of the second bar with space and time coordinates denoted by $\bar{x}$ and $\bar{t}$, which we may therefore take as

$$a\bar{u}(\bar{x}, \bar{t}) + b$$

(using the hint), where a and b are arbitrary constants. To determine these we note that at $t = \bar{t} = 0$ we have

$$u(\tfrac{1}{2}l, 0) = T_0,$$

$$a\bar{u}(\tfrac{1}{2}\bar{l}, 0) + b = \bar{T}_0,$$

and at $x = \bar{x} = 0$ we have

$$u(0, t) = T_1 \qquad t \geq 0,$$

$$a\bar{u}(0, \bar{t}) + b = \bar{T}_1 \qquad \bar{t} \geq 0.$$

Since $\bar{u}(\bar{x}, \bar{t}) = u(x, t)$, we obtain

$$\bar{T}_0 = aT_0 + b \quad \text{and} \quad \bar{T}_1 = aT_1 + b,$$

giving

$$a = \frac{\bar{T}_0 - \bar{T}_1}{T_0 - T_1}, \qquad b = \frac{T_1\bar{T}_0 - \bar{T}_1 T_0}{T_1 - T_0}.$$

The temperature of the centre of the second bar at time t is given by

$$a\bar{u}(\tfrac{1}{2}\bar{l}, t) + b = af(tl^2/\bar{l}^2) + b$$

$$= \frac{(\bar{T}_0 - \bar{T}_1)f(tl^2/\bar{l}^2) + (T_0\bar{T}_1 - \bar{T}_0 T_1)}{T_0 - T_1}.$$

Solution to SAQ 20

Consider the transformation $x = \alpha\bar{x}$, $t = \beta\bar{t}$, under which the equation becomes

$$\frac{1}{\beta^2}\frac{\partial^2\bar{u}}{\partial\bar{t}^2} - \frac{a}{\beta}\frac{\partial\bar{u}}{\partial\bar{t}} - \frac{c^2}{\alpha^2}\frac{\partial^2\bar{u}}{\partial\bar{x}^2} = 0,$$

where $\bar{u}(\bar{x}, \bar{t}) = u(\alpha\bar{x}, \beta\bar{t})$. Thus on choosing $\beta = 1/a$ and $c^2\beta^2/\alpha^2 = 1$, or $\alpha = c/a$ we obtain the required result.

Solution to SAQ 21

We require the solution of

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad 0 < x < 1, \quad t > 0,$$

with the conditions

$$u(0, t) = u(1, t) = 0 \qquad t > 0,$$

$$u(x, 0) = \begin{cases} 2x & 0 \leq x \leq \frac{1}{2} \\ 2(1 - x) & \frac{1}{2} \leq x \leq 1. \end{cases}$$

We consider a solution of the form $u(x, t) = X(x)T(t)$, which on substitution into the heat equation gives

$$\frac{1}{T}\frac{dT}{dt} = \frac{1}{X}\frac{d^2X}{dx^2} = -\alpha^2,$$

where the constant $-\alpha^2$ has been taken to be negative for reasons which will be clear shortly.

The solutions are then

$$T(t) = Ae^{-\alpha^2 t},$$

$$X(x) = B\sin\alpha x + C\cos\alpha x.$$

The boundary conditions at $x = 0, 1$ give:

$$X(0) = C = 0$$

$$X(1) = B\sin\alpha = 0$$

so that $\alpha = n\pi (n = 1, 2, \ldots)$ and the general solution is

$$u(x, t) = \sum_{n=1}^{\infty} A_n \sin n\pi x e^{-n^2\pi^2 t}.$$

(If we had not chosen $-\alpha^2$ as the constant, we would have obtained the trivial solution $u \equiv 0$.) In particular when $t = 0$

$$u(x, 0) = \sum_{n=1}^{\infty} A_n \sin n\pi x.$$

But for $0 \leqslant x \leqslant 1$, $u(x, 0)$ is given, and we have

$$A_n = 2\int_0^1 u(x, 0) \sin n\pi x \, dx$$

$$= 4\int_0^{\frac{1}{2}} x \sin n\pi x \, dx + 4\int_{\frac{1}{2}}^1 (1 - x) \sin n\pi x \, dx.$$

Substituting $1 - x$ for x in the second integral we find that

$$A_n = 4(1 - (-1)^n)\int_0^{\frac{1}{2}} x \sin n\pi x \, dx,$$

since $\cos n\pi = (-1)^n$. It follows that if n is even $A_n = 0$.

The integral can be integrated by parts to give

$$A_{2n+1} = \frac{8 \sin[(2n + 1)\pi/2]}{\pi^2(2n + 1)^2}$$

and so

$$u(x, t) = \frac{8}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n + 1)^2} \sin \frac{(2n + 1)\pi}{2} \sin(2n + 1)\pi x \exp(-(2n + 1)^2\pi^2 t),$$

which has the required convergence properties since $x \longmapsto u(x, 0)$ is continuous, $u(0, 0) = u(1, 0) = 0$, and $\int_0^{\frac{1}{2}} 4dx + \int_{\frac{1}{2}}^1 4dx$ is finite.

Solution to SAQ 22

If x measures the distance around the circumference, the heat equation is approximately (for a *thin* ring)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

where the conductivity has been put equal to unity, and $u(x, t)$ gives the temperature. When $x = 2l$ we return to the point $x = 0$, so $u(x, t)$ must be periodic in x with period $2l$:

$$u(x, t) = u(x + 2l, t) \qquad t \geqslant 0.$$

As in the solution to SAQ 21 we find that a separated solution is given by

$$e^{-\alpha^2 t}(A \sin \alpha x + B \cos \alpha x)$$

and since this must be periodic in x, it is necessary that

$$\alpha = \frac{n\pi}{l}.$$

Thus the general solution is

$$u(x, t) = A_0 + \sum_{n=1}^{\infty} \left(A_n \cos \frac{n\pi}{l} x + B_n \sin \frac{n\pi x}{l}\right) \exp\left[-\left(\frac{n\pi}{l}\right)^2 t\right]$$

and at time $t = 0$ this is

$$u(x, 0) = A_0 + \sum_{n=1}^{\infty} \left(A_n \cos \frac{n\pi x}{l} + B_n \sin \frac{n\pi x}{l} \right)$$

$$= \sin^2 \frac{\pi x}{2l}$$

$$= \frac{1}{2}\left(1 - \cos \frac{\pi x}{l}\right).$$

By equating coefficients we obtain

$$u(x, t) = \frac{1}{2}\left(1 - e^{-(\pi/l)^2 t} \cos \frac{\pi x}{l}\right).$$

Solution to SAQ 23

Using the method of separation of variables and writing $u(x, y) = X(x)Y(y)$ the equation becomes

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{X}\frac{dX}{dx} + \frac{1}{Y}\frac{d^2Y}{dy^2} = 0,$$

with

$$Y(0) = Y(\pi) = 0$$

and

$$X(0) = 0.$$

As usual we write

$$X'' + X' - \lambda X = 0$$

and

$$Y'' + \lambda Y = 0.$$

The boundary conditions on Y yield the eigenvalues

$$\lambda = n^2 \qquad n \in Z^+,$$

and

$$Y_n(y) = \sin ny.$$

Let X_n be the general solution of

$$X_n'' + X_n' - n^2 X_n = 0.$$

Then we set

$$u(x, y) = \sum_{n=1}^{\infty} X_n(x) \sin ny;$$

the boundary condition

$$u(\pi, y) = \sin y$$

gives, formally,

$$\sin y = \sum_{n=1}^{\infty} X_n(\pi) \sin ny,$$

where the Fourier coefficients come out as

$$X_1(\pi) = 1, \quad X_n(\pi) = 0 \qquad n > 1.$$

Solving the boundary value problems for X_n we obtain

$$X_1(x) = e^{(\pi - x)/2} \frac{\sinh(\sqrt{5}x/2)}{\sinh(\sqrt{5}\pi/2)}$$

$$X_n \equiv 0 \qquad n > 0.$$

Thus,

$$u(x, y) = e^{(\pi - x)/2} \frac{\sinh(\sqrt{5}x/2)}{\sinh(\sqrt{5}\pi/2)} \sin y.$$

The validity of the solution is guaranteed, since it is clearly continuous on $[0, \pi] \times [0, \pi]$ and sufficiently differentiable in $(0, \pi) \times (0, \pi)$.

Solution to SAQ 24

In Equation (16.4), *W: page 73*, we set $\rho \equiv 1$, $\phi_n(x) = \sin nx$, $a = 0$, $b = \pi$; thus we we have, since $\{\sin nx : n \in Z^+\}$ is complete on $[0, \pi]$,

$$\lim_{n \to \infty} \frac{\left(\int_0^\pi f(x) \sin nx \, dx\right)^2}{\int_0^\pi \sin^2 nx \, dx} = 0$$

if $\int_0^\pi f(x)^2 \, dx$ is finite. Putting $f(x) = x^\alpha$ we find that

$$\int_0^\pi f(x)^2 \, dx = \left[\frac{x^{2\alpha+1}}{2\alpha + 1}\right]_0^\pi,$$

which is finite if $\alpha > -\frac{1}{2}$, and since

$$\int_0^\pi \sin^2 x \, dx = \frac{\pi}{2}$$

it is evident that

$$\lim_{n \to \infty} \int_0^\pi x^\alpha \sin nx \, dx = 0 \qquad \text{for } \alpha > -\tfrac{1}{2}.$$

Solution to SAQ 25

Since f is odd its Fourier series may be expressed in the form

$$f(x) \sim \sum_{n=1}^{\infty} b_n \sin nx,$$

where

$$b_n = \frac{2}{\pi} \int_0^\pi \sin nx \cos x \, dx.$$

This may be written in the form

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_0^\pi (\sin(n+1)x + \sin(n-1)x) \, dx \\ &= -\frac{1}{\pi} \left[\frac{\cos(n+1)x}{n+1} + \frac{\cos(n-1)x}{n-1}\right]_0^\pi \qquad \text{if } n \neq 1 \\ &= -\frac{1}{\pi}((-1)^{n+1} - 1)\left(\frac{1}{n+1} + \frac{1}{n-1}\right). \end{aligned}$$

Thus $b_n = 0$ if n is odd ($\neq 1$) and if $n = 2m$

$$b_n = \frac{8m}{\pi(4m^2 - 1)}.$$

Also,

$$b_1 = \frac{2}{\pi} \int_0^\pi \sin x \cos x \, dx = \frac{2}{\pi}\left[\frac{\sin^2 x}{2}\right]_0^\pi = 0.$$

Thus

$$f(x) \sim \frac{8}{\pi} \sum_{m=1}^{\infty} \frac{m \sin 2mx}{(4m^2 - 1)}.$$

Solution to SAQ 26

Since f is even its Fourier series will contain only cosines, and the coefficients will be

$$a_n = \frac{2}{\pi}\int_0^{\pi} x^2 \cos nx\, dx \qquad n = 1, 2, \ldots$$

$$= \frac{2}{\pi}\,\mathrm{Re}\int_0^{\pi} x^2 e^{inx}\, dx,$$

where Re means the real part of. The integral is best integrated by parts:

$$\int_0^{\pi} x^2 e^{inx}\, dx = \left[\frac{x^2 e^{inx}}{in} + \frac{2xe^{inx}}{n^2} - \frac{2e^{inx}}{in^3}\right]_0^{\pi}$$

$$= \pi^2\frac{(-1)^n}{in} + \frac{2\pi}{n^2}(-1)^n - \frac{2}{in^3}[(-1)^n - 1],$$

so that

$$a_n = \frac{4}{n^2}(-1)^n = \frac{4\cos n\pi}{n^2}.$$

Also,

$$a_0 = \frac{2}{\pi}\int_0^{\pi} x^2\, dx = \frac{2\pi^2}{3}.$$

The Fourier series converges to $f(x)$ at each point, so that

$$f(x) = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}\frac{\cos n\pi \cos nx}{n^2}.$$

Since $f(\pi) = f(-\pi)$ the Fourier series converges to π^2 at $x = \pi$, and we have

$$\pi^2 = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}\frac{\cos^2 n\pi}{n^2}$$

i.e.

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Solution to SAQ 27

The derived function of f is

$$f' : x \longmapsto \alpha(-x)^{-\alpha-1} \qquad x \in (-\pi, 0)$$

$$\qquad -\alpha x^{-\alpha-1} \qquad x \in (0, \pi).$$

As $x \to 0^+$ or 0^-, f' is unbounded since $\alpha > 0$ and so the limits

$$f'(0+0) \quad \text{and} \quad f'(0-0)$$

do not exist, i.e. f is not piecewise smooth.

Nevertheless, we have seen in *W: page 79* that it is sufficient if Dini's test is satisfied; this will be the case at each point where f is differentiable, provided f is absolutely integrable. Now

$$\int_{-\pi}^{\pi} |f(t)|\, dt = \int_{-\pi}^{\pi} |t|^{-\alpha}\, dt$$

$$= 2\lim_{\varepsilon \to 0^+}\int_{\varepsilon}^{\pi} t^{-\alpha}\, dt \quad \text{since } |f| \text{ is an even function}$$

$$= 2\lim_{\varepsilon \to 0^+}\left[\frac{t^{-\alpha+1}}{1-\alpha}\right]_{\varepsilon}^{\pi}$$

$$= \frac{2}{1-\alpha}\pi^{1-\alpha} \quad \text{provided } 0 < \alpha < 1.$$

Thus, if $0 < \alpha < 1$ then the Fourier series of f converges to f at each point in $(-\pi, 0) \cup (0, \pi)$.

Unit 7 Motion of Overhead Electric Train Wires

Contents

Set Books

G. D. Smith, *Numerical Solution of Partial Differentiation Equations* (Oxford, 1971).

H. F. Weinberger, *A First Course in Partial Differential Equations* (Blaisdell, 1965).

It is essential to have these books; the course is based on them and will not make sense without them. They are referred to in the text as S and W respectively.

Unit 7 is not based on either set book.

Conventions

Before working through this text make sure you have read *A Guide to the Course: Partial Differential Equations of Applied Mathematics*. References to Open University courses in mathematics take the form:

> *Unit M100 13, Integration II* for the Mathematics Foundation Course.
>
> *Unit M201 23, The Wave Equation* for the Linear Mathematics Course.

Acknowledgements

We are grateful for assistance given by Mr. E. A. Cardwell and Mr. H. H. Ogilvy of British Rail in the preparation of material included in this text. Some of the theory is due to G. Gilbert and H. E. H. Davies, "Pantograph Motion on a Nearly Uniform Railway Overhead Line" in *Proceedings of the Institute of Electrical Engineers*, 1966, Vol. 113, No. 3.

The photograph in the Introduction is reproduced by permission of London Midland Region (B.R.).

7.0 INTRODUCTION

This unit is essentially a case study using some of the techniques developed earlier in the course. In it we investigate the modelling of the behaviour of an overhead wire when a high-speed electric train passes. We start with a brief history of the development of railways and their electrification.

7.0.1 History of the Electric Train

The earliest railways in Britain are dated at about 1550; by the end of the sixteenth century horse-drawn trains were common, particularly in the North-East coal mining area, and by the eighteenth century horse-drawn trains were found all over Wales, Scotland and England. The first steam locomotive, built by Trevithick, ran in 1804 but was too heavy for the track which kept breaking; and in 1823 Stephenson built the first railway for carrying both freight and passengers.

Efforts at using battery-operated electric locomotives date from 1835, but the first successful demonstration was not until 1879, at an exhibition in Berlin. Following this, the first public electric railway was opened near Berlin in 1881, and two years later another was opened in Brighton. The use of electricity for main-line trains, however, did not occur until the beginning of this century. By 1920 most European countries had at least a small section of electrified track.

The advantages of electric traction are diverse. Electric locomotives are quieter, less polluting, can accelerate faster, and are easier and cheaper to maintain than other locomotives. In addition, it is generally recognized that if cheap electricity is available and if there is sufficient traffic, this is the most economical and efficient method of traction. Part of the reason for this is that an electric locomotive converts power rather than generates it, and draws its energy from a central power station. This means that resources can be used more efficiently, and that for short times a locomotive can develop power greatly in excess of its nominal rating in order to start or climb hills. (For example, a locomotive nominally rated at 4000 h.p. or 3 MW* has been observed to develop 10 000 h.p. (7.5 MW); this would not be possible with either a steam or a diesel locomotive.) Against this, however, is the high capital cost of electrification.

There are two main types of system: alternating current (a.c.) and direct current (d.c.). In the early days d.c. systems were more popular and operated at either 1500 or 3000 volts, although a 600 volt system operates in south-east England. D.c. systems need frequent substations, and the overhead wire, or third rail, needs to be large and heavy giving rise to a large capital cost. Modern systems use a.c. at 25 000 volts; this requires fewer substations and thinner overhead wire, and is consequently cheaper than the lower voltage d.c. systems. The first extensive use of a.c. was in north-east France, which was converted in 1952; subsequently Argentina, Belgian Congo (now Zaire), Britain, China, India, Japan and Portugal built similar systems.

In order to reduce maintenance and equipment costs it is necessary to have, amongst other things, an understanding of the dynamic properties of the overhead wire and the pantograph system. In this unit we describe one set of approximations which have been made with a view to gaining a theoretical understanding of the motion of this complicated system. The approximations described have the advantage that they lead to a mathematical model to which analytical solutions are possible: to achieve this the approximations made are quite severe. However, the solutions obtained give a reasonable qualitative idea of the wire motion. To obtain a better description of the motion it is necessary to make fewer approximations; if we did this analytical solutions could no longer be obtained, and numerical methods would be necessary.

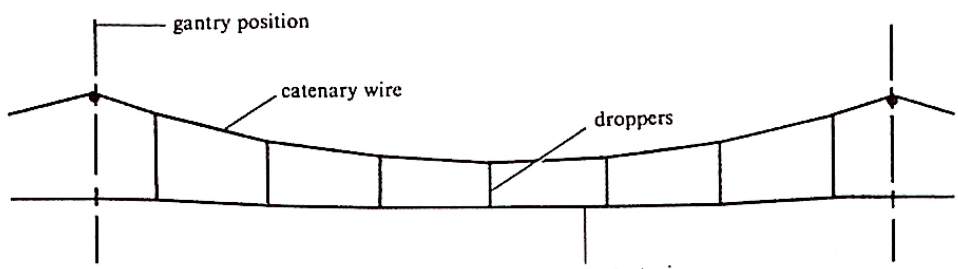
* megawatts.

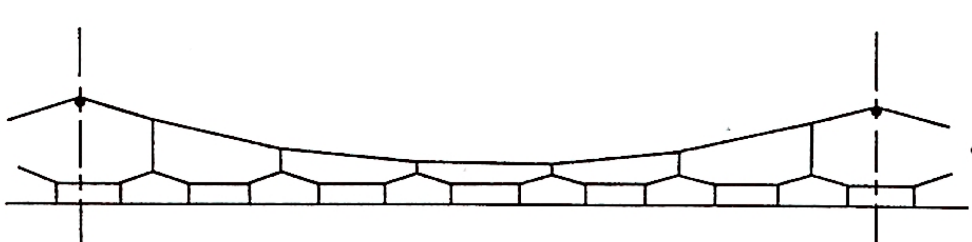
Before embarking on the mathematics we describe the salient features of the system that our equations purport to describe.

7.0.2 The Overhead Equipment

A variety of overhead systems is in use and three different kinds are shown. The *trolley* wire is the simplest system of all but is suitable for low speeds only. The *sagged simple* wire ("sagged" because of a slight but important sag of about 10 cm —about 4″—in the bottom wire) is suitable for moderately high speeds. The *compound* wire is used on some high-speed lines in Britain.







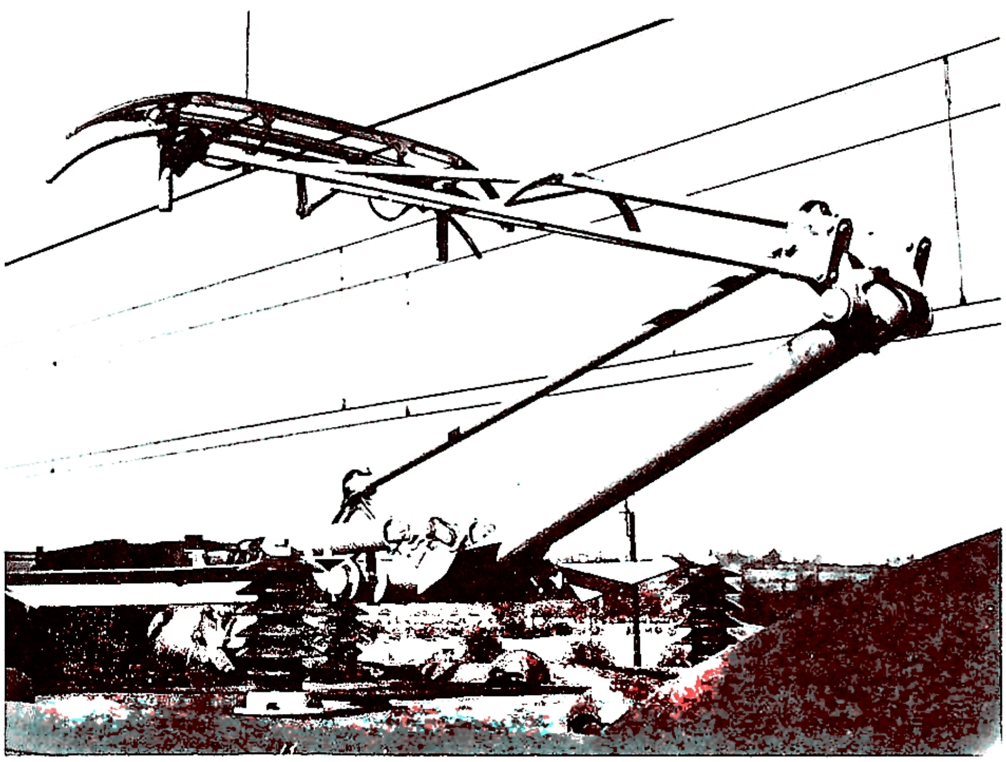
An important property of the wire is its **compliance**; this is the distance by which a unit force lifts the wire. It varies from point to point, being least at gantries and supports. The **stiffness** of the wire is the reciprocal of its compliance.

Owing to the variation in compliance, a pantograph moving at a constant low speed along a wire and exerting a constant force on the wire oscillates with frequency v/l, where v is the train's speed and l is the spacing between the supports, and with an amplitude depending upon the variation in compliance of the contact wire. At high speeds this motion becomes extremely complex and is determined by properties of the wire and the pantograph; it is this motion that we are trying to understand.

The contact wire, together with its supporting equipment, is quite a complicated physical system. Since it is supported at approximately equal intervals (about 60 m,

or 200 ft) its physical properties are periodic along its length and it can oscillate at its own natural frequencies; it is practically undamped so that disturbances propagate large distances from their source; this has important consequences for trains with two or more pantographs.

In this unit we shall describe a model of the wire which is sufficiently simple to provide analytic solutions, but sufficiently accurate to give a qualitative idea of the motion.



The pantograph is the device which conducts electricity from the wire to the motors; the current passes through a simple collector strip on the pantograph sliding along the wire. Although this may sound simple, the system is complicated by several factors.

The first complication is caused by the variation in the height of the wire above the rails, which can be from 6.3 m (20′ 6″) at level crossings to 4.1 m (13′ 5″) at bridges. Throughout this range the static force of the pantograph (i.e. when the train is stationary) must remain almost constant (in practice 90 ± 9 N*). To achieve this the pantograph has to be a fairly complex structure, but because it must respond readily to changes in height it must have a small mass. It also needs to be a quite rugged structure since sideways accelerations due to the motion of the train can be quite substantial. In addition, the profile of the pantograph is important, since at 160 km/hr (100 m.p.h.) the increased vertical force on the wire due to aerodynamic effects on the pantograph could be as much as 22 N. This force depends strongly upon the pantograph design, and British Rail pantographs are designed to keep it small. There are also problems with the formation of ice on the pantograph during winter; the weight of ice decreases the upward force, so altering the characteristic of the pantograph.

Many of these problems arise because of the imposed constraint that the force on the wire must remain within fairly narrow bounds during the motion of the train. Two reasons for this are as follows.

If the contact wire is lifted too far from its equilibrium position the pantograph could strike the supports, possibly causing a great deal of damage. The limits of the lift are in general 15 cm, but 5 cm at bridges.

The force should remain positive, otherwise the pantograph separates from the contact wire and arcing occurs which causes wear on both surfaces. In practice it seems impossible to stop this completely, but the aim is to minimize it.

*newtons.

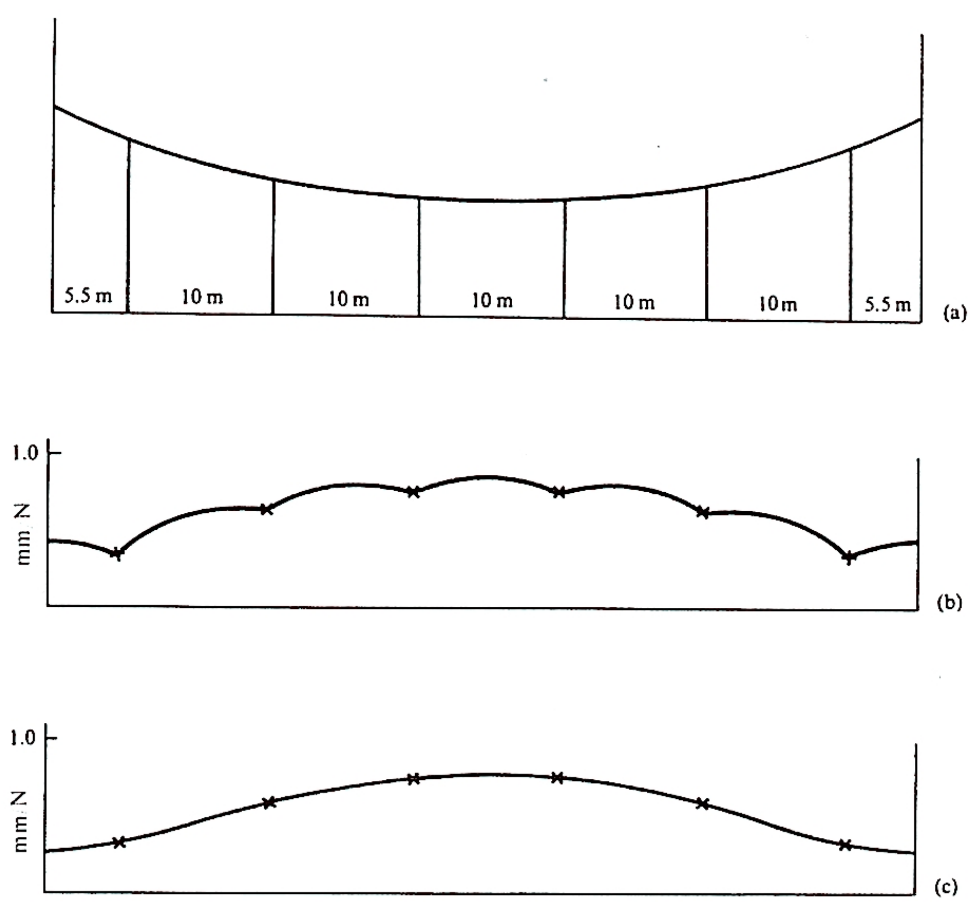
From this brief description it is clear that the pantograph is quite a complicated piece of equipment with its own natural modes of oscillation. In a more detailed model we should take into account the effects of the pantograph's motion: we shall not do this here, but will represent the pantograph by a constant force acting upwards at a point moving along the wire. Besides ignoring the motion of the pantograph we assume, too, that the wire moves only in a vertical plane; clearly there can be motion in the horizontal plane, but in general its omission does not impair the usefulness of the model we will use.

(For those of you interested in the practicalities, there are three horizontal features. The first is the lateral displacement of the wire in a strong wind, which must always be small enough to ensure that the wire does not come off the pantograph; this sometimes determines the span length between the gantries. The second is that the contact wire is arranged to zig-zag from side to side between supports, giving almost uniform wear over the whole pantograph. Finally, the catenary wire need not be vertically above the contact wire, so that the supporting wires are not necessarily vertical. These effects are ignored in our model.)

7.1 A MODEL OF THE WIRE

We now proceed to obtain an approximation to the motion of the wire. We assume that the wire may be represented by a perfectly flexible stretched string, the equation of motion of which was derived in *Unit 1, The Wave Equation* for the case in which there are no body forces.

The complexities introduced by the present problem are due entirely to the forces acting upon the wire. Firstly, we have the pantograph which, as has already been mentioned, we shall represent by an upward force of constant magnitude P_0 acting at a point moving along the wire with constant speed v; for the present we represent this by a force per unit length $F(x, t)$ where x is the distance along the wire and t is the time. Secondly, there are damping forces which we shall assume, fairly accurately, to be proportional to the vertical velocity of the contact wire; these forces (for example, air resistance) are represented by a term of the form $-\eta \partial y/\partial t$ per unit length where η is a positive constant and $y(x, t)$ is the vertical displacement. Thirdly, there is the weight of the wire which is simply $-\rho g$ per unit length where ρ is the line density and g the acceleration due to gravity. Finally, we make our most severe approximation and assume that the wire is continuously supported by an elastic medium the properties of which vary only slightly along the length of the wire.



The diagram shows the variation in compliance between supports for a simple system. In (a) we show the catenary wires with droppers spaced at 10 m intervals. In (b) we have plotted the values of the compliance for a real system; these are given in mm per newton. Notice that the effect of the droppers is significant. In (c) we have "smootned out" the effect of the droppers; this is an approximation made in the subsequent analysis of this unit.

This last approximation needs some explanation. In practice the wire is supported at a discrete number of points, at each of which the restoring force on the wire is proportional to the displacement from its equilibrium position; for a given displacement these forces vary along the wire as the reciprocal of the *compliance* (shown in the figure). It is possible to represent such supports mathematically but analytical solutions of the ensuing equations seem to be impossible. Here we replace these discrete supports by a continuous one; it is supposed that at each point x of the wire

the restoring force per unit length exerted on the wire by the support is proportional to $y(x, t)$. (We can visualize this by imagining that the wire is stuck onto a sheet of rubber.) Since the stiffness of the wire is not constant we represent this force by $-S(x)y(x, t)$, where $S(x)$ is the elasticity/unit length of the supporting medium, and is a continuously differentiable function of x.

The equation representing forced wave motion is derived as for the vibrating string problem in *Unit 1*. In our case we obtain

$$-T\frac{\partial^2 y}{\partial x^2}(x, t) + \rho\frac{\partial^2 y}{\partial t^2}(x, t) = F(x, t) - \eta\frac{\partial y}{\partial t}(x, t) - \rho g - S(x)y(x, t) \qquad (1)$$

where the tension T and the line density ρ are both assumed constant. Some typical values of the constants appearing in this equation, and some other relevant data are given in the table.

tension	T	8900 N
line density	ρ	0.952 kg m^{-1}
mean elasticity/unit length	S_0	1700 N m^{-2}
distance between gantries	l	61 m
damping constant	η	$0.1 \text{ kg m}^{-1}\text{ s}^{-1}$
acceleration due to gravity	g	9.81 m s^{-2}

SAQ 1

Consider a static wire (i.e. y is a function of x only so that $\partial y/\partial t \equiv 0$) with no forcing term ($F \equiv 0$), embedded in an elastic medium of constant elasticity $S(x) = S_0$ = constant and supported at $x = 0$ and $x = l$, so that $y(0, t) = y(l, t) = 0$.

(a) Show that the wire has a shape given by the graph of

$$y(x, t) = \frac{g}{v^2c^2}\left[\frac{\sinh v(l - x) + \sinh vx - \sinh vl}{\sinh vl}\right]$$

where $v = (S_0/T)^{\frac{1}{2}}$, $c = (T/\rho)^{\frac{1}{2}}$.

(b) Show that the maximum displacement is at the mid point and is given by

$$-\frac{g}{v^2c^2}\left[1 - \frac{2\sinh(vl/2)}{\sinh vl}\right].$$

(c) Show that with the data given in the table the maximum sag is about 0.55 cm. (This is in fact less than the diameter of the wire.)

(Solution on p. 23.)

This example demonstrates that the weight of the wire does not significantly alter the shape of the wire in comparison with the uplift (several cms) due to the pantograph. For this reason we may ignore the term ρg in Equation (1) in our modelling.

From the table we see that the damping constant η is small and in the following analysis we shall ignore $\eta(\partial y/\partial t)$ (typical value 0.15 kg s^{-2} compared with a typical value of 50 kg s^{-2} for yS). We do this mainly for convenience in that the ensuing algebra is less tedious: analytical solutions can be obtained without this assumption by exactly the same methods as outlined in the rest of this unit.

[In some cases, although numerically small, this damping term can have a large effect on the solution. In *Unit M201 9, Homogeneous Equations* it was shown that the steady-state solution of the equation

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = E\sin\omega t,$$

which is the one-dimensional analogue of Equation (1), is

$$q = \frac{-E\cos(\omega t - \alpha)}{\omega\left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{\frac{1}{2}}}$$

where

$$\tan\alpha = \frac{1}{R}\left(L\omega - \frac{1}{C\omega}\right).$$

If $R \neq 0$ (equivalent to $\eta \neq 0$ in Equation (1)) the denominator is never zero and the solution is always bounded. But if $R = 0$ the denominator tends to zero when ω^2 approaches $1/LC$ and the solution is unbounded. Thus for small R there are certain *resonant* frequencies for which the amplitude of the solution alters drastically. We shall see later that exactly the same phenomenon occurs in our model.]

With these simplifications Equation (1) may be written as

$$-\frac{\partial^2 y}{\partial x^2} + \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} + \frac{S}{T}y = \frac{F}{T}, \tag{2}$$

where $c = (T/\rho)^{\frac{1}{2}}$. The elasticity $S(x)$ of the supporting medium is periodic with period l (the distance between the supports), so that

$$S(x + l) = S(x).$$

Since we are assuming that the stiffness of the contact wire is nearly uniform, $S(x)$ varies very little between spans and we write

$$S(x) = S_0[1 - \varepsilon f(x)] \tag{3}$$

where ε is a small constant, S_0 is the mean elasticity/unit length and $S_0\varepsilon f(x)$ is a perturbation about this mean.

SAQ 2

(a) Show that $f(x)$ must be periodic with period l.

(b) Since $f(x)$ is periodic what is a convenient way of representing it?

(c) Since the means value of $f(x)$ is zero, what can you deduce about the coefficients in this representation?

(Solution on p. 24.)

One final point about the assumption that ε is small. Although it is a reasonable approximation it is not necessarily true. However, if the approximation were not made we would need to solve the equation by a numerical method. In this event the analytical solution would provide a means of testing the solutions obtained as the value of ε is decreased.

7.2 SOLUTIONS FOR A UNIFORM SUPPORT

7.2.1 A Free Wire

Before considering the effect of the pantograph on the wire it is instructive to consider the free oscillations of the wire in the absence of external forces. The equation of motion is then

$$-\frac{\partial^2 y}{\partial x^2}(x,t) + \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2}(x,t) + v^2[1 - \varepsilon f(x)]y(x,t) = 0, \qquad (4)$$

where we have put $v = (S_0/T)^{\frac{1}{2}}$. General solutions may be obtained only for $\varepsilon = 0$, but since ε is assumed small this is a satisfactory first approximation.

SAQ 3

What does $\varepsilon = 0$ represent physically?

(Solution on p. 24.)

We are interested in wave motion along the wire and we know that for a disturbance moving along an ordinary string (Equation (4) with $v = 0$) this is of the form $y(x,t) = g(x \pm ct)$ as we saw in *Unit 1*, Section 1.2. Such motion can, of course, be analyzed into Fourier components.

Now, in the discussion of the complex Fourier series in *Unit 6, Fourier Series* (Section 6.3) we saw that complex exponentials are more convenient to handle than sines and cosines. We therefore try

$$y(x,t) = \exp i(kx - \omega t) \text{ for some } \omega \in R; \qquad (5)$$

if $k > 0$ the real and imaginary parts represent wave motion in the positive x-direction, with **angular frequency** ω, **wave number** k, and **(phase) velocity** ω/k. (If this heuristic discussion of the solution is too woolly for your liking, you might care to solve (4) by separation of variables with the condition that $y(x,t)$ is bounded for all t, and so derive (5).)

SAQ 4

Substitute the complex solution represented by Equation (5) into Equation (4) with $\varepsilon = 0$ to show that solutions of this form exist provided the frequency ω and the wave number k are related by

$$k = (\omega^2/c^2 - v^2)^{\frac{1}{2}},$$

and that the phase velocity is given by

$$c(1 - v^2c^2/\omega^2)^{-\frac{1}{2}}.$$

(Solution on p. 24.)

We see from SAQ 4 that when $v \neq 0$ a wave represented by the real or imaginary part of $\exp i(kx - \omega t)$ has a velocity which depends upon frequency. This is different from the simple systems considered in *Unit 1*. We notice that for large frequencies ($\omega \gg cv$) we have $k \simeq \omega/c$; in this case the wire is vibrating at too high a frequency for the supporting medium to have effect, and the wire behaves as though the medium were not present.

There is a *cut-off frequency* $\omega_c = cv$ below which waves will not propagate. For $\omega < cv$, Equation (5) does not yield a real k and no wave motion is possible. When $\omega = \omega_c$ the wave length is infinite and the whole wire moves bodily up and down at this frequency, which is the fundamental frequency of the supporting medium.

SAQ 5

Consider an infinite wire along the x-axis. The wire is embedded in an elastic medium with constant elasticity, so that its equation of motion is given by Equation (4) with $\varepsilon = 0$. At the origin the wire is forced to vibrate with angular frequency Ω, so that

$$y(0,t) = \cos \Omega t [= \mathrm{Re}(e^{-i\Omega t})] \qquad t \in R.$$

Describe the motion when $\Omega > cv$, $\Omega = cv$ and $\Omega < cv$ assuming that it is completely symmetric about the origin.

HINT: Determine an appropriate complex solution and take the real part at the end of the calculation.

(Solution on p. 25.)

In physics it is common for the velocity of a wave to depend upon its frequency. This phenomenon is called **dispersion** and is important in electromagnetic theory, optics and other areas of physics.

7.2.2 A Constant Point Force

We now introduce into our model the effect of the pantograph, represented by a constant point force moving along the wire at a constant speed.

In this case it is convenient to rewrite the equation of motion by transforming co-ordinates from the fixed frame of reference to a moving frame. In the next SAQ you see that there is one such transformation that leaves the wave equation invariant, although this is not the one we shall use in practice.

SAQ 6

Show that the wave equation

$$\frac{\partial^2 y}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} = 0$$

is invariant (i.e. not changed in form) under the coordinate transformation

$$\bar{x} = \frac{x - ut}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}, \quad \bar{t} = \frac{t - \frac{u}{c^2}x}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}},$$

where u and c are constants with $u < c$.

(This SAQ involves the chain rule and some lengthy manipulation.)

(Solution on p. 25.)

SAQ 7

(i) An infinite wire lies along the x-axis. At the origin it is forced to move so that its displacement is given by $y(0, t) = f(t)$ where f is a known function. Assuming that the disturbance travels away from the source show that the motion is represented by

$$y(x, t) = f\left(\frac{ct - x}{c}\right) \qquad x > 0,\ t \in R,$$

$$y(-x, t) = y(x, t) \qquad x \in R, \qquad t \in R,$$

where c^2 is the ratio of the tension T in the wire to its line density.

(ii) Find the force exerted by the tension in the wire at the origin, assuming that the gradient of the wire is small there. Express your result in terms of f.

(Solution on p. 26.)

SAQ 8

A small ring, moving along an infinite straight wire with constant velocity u, is forced to move so that its displacement perpendicular to the equilibrium position of the wire at time t is given by $a \sin \omega t$. By using results obtained in the solutions of

the previous two SAQs determine the motion of the wire caused by the ring in the case $u < c$, where c is defined as in SAQ 7.

(Solution on p. 27.)

For our model of the idealized pantograph, which acts at a point on the wire with a constant force P_0, we transform to the coordinate system given by

$$\bar{x} = x - vt, \qquad \bar{t} = t,$$

where v is the (constant) speed of the train (and hence of the pantograph). Note that for all $\bar{t}$ the pantograph is at the origin $\bar{x} = 0$.

The equation of motion for the wire is, in the original coordinate system,

$$-\frac{\partial^2 y}{\partial x^2}(x,t) + \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2}(x,t) + v^2[1 - \varepsilon f(x)]y(x,t) = 0 \qquad x \neq vt.$$

The situation at $x = vt$ involves the constant point force P_0, which cannot be included directly in the differential equation, in which we can consider only forces per unit length. We shall take it into account in the boundary conditions.

SAQ 9

Show that, in the new coordinate system, the equation of motion becomes

$$-(1 - v^2/c^2)\frac{\partial^2 \bar{y}}{\partial \bar{x}^2}(\bar{x},\bar{t}) - \frac{2v}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{x}\,\partial \bar{t}}(\bar{x},\bar{t}) + \frac{1}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{t}^2}(\bar{x},\bar{t})$$
$$+ v^2[1 - \varepsilon f(\bar{x} + v\bar{t})]\,\bar{y}(\bar{x},\bar{t}) = 0 \qquad (6)$$

for $\bar{x} \neq 0$, where

$$\bar{y}(\bar{x},\bar{t}) = y(\bar{x} + v\bar{t}, \bar{t}).$$

(Solution on p. 28.)

In this section we are considering only a uniform support ($\varepsilon = 0$). In this case the pantograph senses the same situation at all times and consequently we might expect that in the frame fixed relative to the pantograph the solution is time independent. Thus with $\varepsilon = 0$ we may put $\partial\bar{y}/\partial\bar{t} = 0$ and Equation (6) reduces to

$$-(1 - v^2/c^2)\frac{\partial^2 \bar{y}}{\partial \bar{x}^2} + v^2\bar{y} = 0 \qquad \bar{x} \neq 0. \qquad (7)$$

Note however that $\partial y/\partial t \neq 0$. In the following SAQ we ask you to derive the solution of Equation (7).

SAQ 10

(a) Show that the general solution to Equation (7) is

$$\bar{y}(\bar{x},\bar{t}) = A_+e^{-\sigma\bar{x}} + B_+e^{\sigma\bar{x}} \qquad \bar{x} > 0,$$
$$\bar{y}(\bar{x},\bar{t}) = A_-e^{-\sigma\bar{x}} + B_-e^{\sigma\bar{x}} \qquad \bar{x} < 0,$$

where $\sigma = v(1 - v^2/c^2)^{-\frac{1}{2}}$, and where $A_\pm$ and $B_\pm$ are constants. Note that as v increases so does σ.

(b) What subsidiary conditions, associated with the problem under investigation, should be applied?

(c) Use the subsidiary conditions to show that the shape of the wire is given by $\bar{y}(\bar{x},\bar{t}) = Ae^{-\sigma|\bar{x}|}$ where A is a constant. (8)

(d) By resolving forces at the origin show that in the case of zero velocity

$$A = \frac{P_0}{2vT} \qquad (9)$$

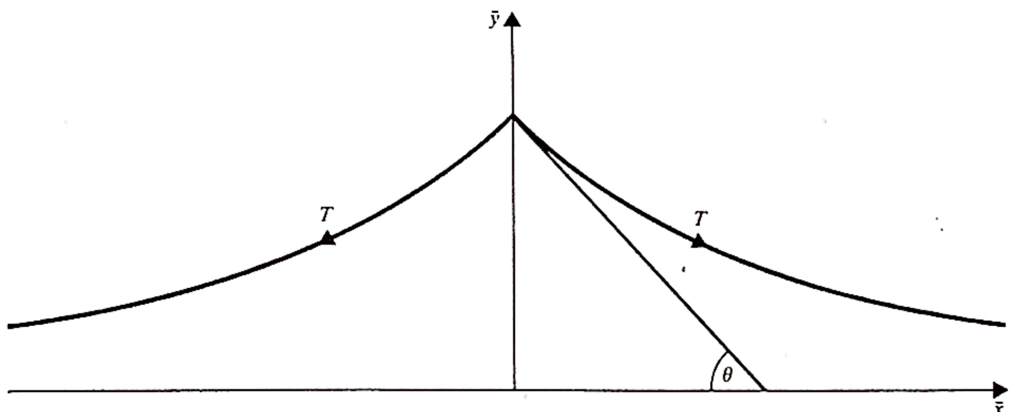
where P_0 is the constant force exerted by the pantograph on the contact wire.

Assume that the gradient of the wire is small at the point of contact.

(Solution on p. 28.)

When the velocity v is not zero the boundary condition at the origin is obtained by resolving forces in the same way as in SAQ 10(d), but now it is necessary to take into account the motion of the wire (since we are using coordinates moving with the pantograph which is effectively stationary with the wire sliding over it).

As the wire slides over the pantograph it accelerates downwards. For, consider a point on the wire: just before the origin its horizontal component of velocity is v and, since it is travelling at an angle θ to the horizontal, its vertical velocity is $v \tan\theta$ (see diagram) and immediately after the origin it has the same component of velocity downwards; the change in velocity is thus $2v \tan\theta$.



In a time Δt a MASS $\rho v \Delta t$ passes the origin and experiences an ACCELERATION $2v \tan\theta/\Delta\bar{t}$ in the downward direction. Also, there is a FORCE $P_0 - 2T\sin\theta$ acting upwards. Thus, on using Newton's Second Law of Motion,

FORCE = MASS × ACCELERATION,

we obtain

$$P_0 - 2T\sin\theta = -\rho v \Delta\bar{t} 2v \sin\theta/\Delta\bar{t}.$$

Hence, since the gradient is small near the origin, $\sin\theta \simeq \tan\theta$ and

$$P_0 = 2T\tan\theta(1 - v^2/c^2) \quad \text{since} \quad c^2 = T/\rho$$

$$= \frac{2Tv^2}{\sigma^2}\tan\theta \qquad \text{since} \quad \sigma^2 = v^2(1 - v^2/c^2)^{-1}.$$

But

$$\tan\theta = \left(\frac{\partial\bar{y}}{\partial\bar{x}}\right)_{\bar{x}=0^-}$$

$$= A\sigma \qquad \text{using Equation (8).}$$

Thus, approximately,

$$A = \frac{P_0\sigma}{2Tv^2}$$

and

$$\bar{y}(\bar{x}, \bar{t}) = \frac{P_0}{2Tv^2}\sigma e^{-\sigma|\bar{x}|}. \tag{10}$$

The shape of this curve is depicted in the figure above.

Since the derivative at the origin is proportional to

$$\frac{\sigma^2}{v^2} = \left(1 - \frac{v^2}{c^2}\right)^{-1},$$

it increases with velocity and the slope of the wire theoretically gets 'steeper'. So the change in the derivative across the origin increases at higher speeds. As this happens the theory becomes less appropriate for two reasons. Firstly, a real wire cannot be allowed to have a discontinuity in its first derivative since it is designed to resist

kinking at the velocities encountered. Secondly, our derivation of the wave equation assumes that all gradients are small, and this is clearly not the case near the origin for large velocities. Away from the origin however, the model remains good.

We note also that our solution is restricted to $v < c$. At $v = c$, the coefficient of $\partial^2\bar{y}/\partial\bar{x}^2$ vanishes, and for $v > c$, σ is no longer real and the nature of the solution is altered.

7.3 SOLUTIONS FOR A NON-UNIFORM SUPPORT

7.3.1 The Equation of Motion

We now investigate the case of a non-uniform support, i.e. $\varepsilon \neq 0$. Writing L for the operator

$$-(1 - v^2/c^2)\frac{\partial^2}{\partial \bar{x}^2} - \frac{2v}{c^2}\frac{\partial^2}{\partial \bar{x}\,\partial \bar{t}} + \frac{1}{c^2}\frac{\partial^2}{\partial \bar{t}^2} + v^2$$

Equation (6) becomes

$$L[\bar{y}](\bar{x}, \bar{t}) = \varepsilon v^2 f(\bar{x} + v\bar{t})\bar{y}(\bar{x}, \bar{t}) \qquad \bar{x} \neq 0, \quad \bar{t} \in R. \tag{11}$$

Since we shall assume that the effects of the non-uniformity of the support are small, we suppose that the solution is of the form

$$\bar{y} = y_1 + \varepsilon y_2 + O(\varepsilon^2)^*$$

where y_1 and y_2 are independent of ε. Substituting this into Equation (11) and rearranging we obtain

$$L[y_1](\bar{x}, \bar{t}) + \varepsilon\{L[y_2](\bar{x}, \bar{t}) - v^2 f(\bar{x} + v\bar{t})y_1(\bar{x}, \bar{t})\}$$
$$- \varepsilon^2 v^2 f(\bar{x} + v\bar{t})y_2(\bar{x}, \bar{t}) = O(\varepsilon^2) \qquad \bar{x} \neq 0.$$

Since this is true for all ε, it is true for $\varepsilon = 0$ with the consequence that

$$L[y_1](\bar{x}, \bar{t}) = 0 \qquad \bar{x} \neq 0,$$

so that y_1 is the uniform support solution given by Equation (10), that is

$$y_1(\bar{x}, \bar{t}) = \frac{P_0}{2Tv^2}\sigma\, e^{-\sigma|\bar{x}|}.$$

It then follows that

$$L[y_2](\bar{x}, \bar{t}) = v^2 f(\bar{x} + v\bar{t})y_1(\bar{x}, \bar{t}) \qquad \bar{x}, \bar{t} \in R, \tag{12}$$

which is a differential equation for y_2. Since f is a periodic function of period l we can express it as a complex Fourier series, assuming that it has the required smoothness property. Thus we write

$$f(z) = \sum_{n=-\infty}^{\infty} a_n \exp iknz \qquad z \in R,$$

where $k = 2\pi/l$, $a_0 = 0$ (see SAQ 2) and $a_n = \bar{a}_{-n}$ for $n > 0$ by SAQ 17 of *Unit 6, Fourier Series*.

Equation (12) is linear and so the analysis can be made simpler by considering each harmonic in turn and adding the results, i.e. we write

$$y_2 = \sum_{n=-\infty}^{\infty} \zeta_n$$

where

$$L[\zeta_n](\bar{x}, \bar{t}) = a_n v^2 \exp ikn(\bar{x} + v\bar{t})y_1(\bar{x}, \bar{t}). \tag{13}$$

The first harmonic in Equation (13) is given by

$$L[\zeta_1](\bar{x}, \bar{t}) = \frac{a_1 P_0 \sigma}{2T}\exp[-\sigma|\bar{x}| + ik(\bar{x} + v\bar{t})] \qquad \bar{x}, \bar{t} \in R, \tag{14}$$

The solution will be the sum of the general solution of the homogeneous equation $L[\zeta] = 0$ and any particular solution of the nonhomogeneous equation (14). Since we expect a wave motion, as in the analogous case of a free wire with a uniform support (Section 7.2.1), we consider solutions of the form

$$\zeta(\bar{x}, \bar{t}) = \exp(\lambda\bar{x} + i\omega\bar{t}),$$

where $\lambda \in C$, $\omega \in R$.

* The O notation was discussed in *Unit 5, Initial Value Problems*.

SAQ 11

Show that $\zeta(\bar{x}, \bar{t}) = \exp(\lambda\bar{x} + i\omega\bar{t})$ satisfies the homogeneous equation $L[\zeta] = 0$ if $\lambda = \lambda_+$ or λ_- where

$$\left(1 - \frac{v^2}{c^2}\right)\lambda_\mp = -i\frac{v\omega}{c^2} \pm v\left[\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{v^2c^2}\right]^{\frac{1}{2}}.$$

Supposing that

$$\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{v^2c^2} > 0 \qquad (15)$$

what are the forms of the solutions for $\bar{x} > 0$ and $\bar{x} < 0$? What forms do the solutions take if

$$\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{v^2c^2} < 0?$$

(Solution on p. 29.)

Equation (14) represents the equation of motion of a vibrating system with a periodic forcing term of frequency kv. In practice, there are damping terms, e.g. air resistance, so that all the modes of oscillation of frequency $\omega \neq kv$ die out (see *Unit M201 11, Nonhomogeneous Equations*, Section 11.1.3) and we need not consider them further.

On putting $\omega = kv$ into the inequality (15) it is clear that we obtain fundamentally different solutions depending upon whether v is greater or less than the critical velocity

$$v_c = \frac{c}{\left(1 + \frac{k^2}{v^2}\right)^{\frac{1}{2}}}.$$

Typically $v_c \simeq 210$ mph. If the train speed is less than this critical speed then the waves die out in front and behind the pantograph. Otherwise $\lambda_\pm$ are purely imaginary and the waves extend (theoretically) to infinity in both directions. We consider only the former case.

To look for a particular solution of Equation (14) we try the form

$$A\frac{a_1P_0\sigma}{2T}\exp[-\sigma|\bar{x}| + ik(\bar{x} + v\bar{t})]$$

which, on substitution into (14), gives

$$A = \frac{1}{k(k + 2i\sigma)} \qquad \bar{x} > 0,$$

$$A = \frac{1}{k(k - 2i\sigma)} \qquad \bar{x} < 0.$$

Combining kernel functions and the particular solution, we obtain the general solution

$$\zeta_1(\bar{x}, \bar{t}) = \frac{a_1P_0\sigma}{2Tk}\left\{\frac{\exp[-\sigma\bar{x} + ik\bar{x}]}{k + 2i\sigma} + C_+e^{\lambda_-\bar{x}}\right\}e^{ikv\bar{t}} \qquad \bar{x} > 0,$$

$$\zeta_1(\bar{x}, \bar{t}) = \frac{a_1P_0\sigma}{2Tk}\left\{\frac{\exp[\sigma\bar{x} + ik\bar{x}]}{k - 2i\sigma} + C_-e^{\lambda_+\bar{x}}\right\}e^{ikv\bar{t}} \qquad \bar{x} < 0. \qquad (16)$$

The constants C_+ and C_- are found by using the conditions that ζ_1 and $\partial\zeta_1/\partial\bar{x}$ are continuous at the origin. The additional condition that $\partial\zeta_1/\partial\bar{x}$ be continuous at $\bar{x} = 0$ may be justified on the grounds that the discontinuity in $\partial\bar{y}/\partial\bar{x}$ at $\bar{x} = 0$ is completely included in y_1. Thus the left-hand side of Equation (12), and *ipso facto* of Equation (14), is defined at $\bar{x} = 0$. In particular, $\partial^2\zeta_1/\partial\bar{x}^2$ is defined there and hence $\partial\zeta_1/\partial\bar{x}$ is continuous. After several lines of manipulation we obtain

$$C_\pm = \frac{2\sigma}{\lambda_+ - \lambda_-}\frac{k + 2i\lambda_\pm}{k^2 + 4\sigma^2}.$$

It is convenient to write the solution in the form

$$\zeta_1(\bar{x},\bar{t}) = \frac{a_1 P_0 \sigma}{2Tkd}\left\{e^{-\sigma|\bar{x}|}(k \mp 2i\sigma)e^{ik(\bar{x}+v\bar{t})}\right.$$

$$\left. + \frac{\sigma}{\sigma_1} e^{-\sigma_1|\bar{x}|}[k(1 + 2\Omega) \pm 2i\sigma_1]e^{-ik(\Omega\bar{x} - v\bar{t})}\right\}. \qquad (17)$$

where the upper (lower) sign is taken for $\bar{x} \geqslant 0$ ($\bar{x} \leqslant 0$).

$$\Omega = \frac{v^2/c^2}{(1 - v^2/c^2)},$$

$$\sigma_1 = \sigma\left[\frac{1 - \left(1 + \frac{k^2}{\nu^2}\right)\frac{v^2}{c^2}}{1 - \frac{v^2}{c^2}}\right]^{\frac{1}{2}} = \sigma\left[1 - \frac{k^2\Omega^2}{\nu^2}\right]$$

and

$$d = k^2 + 4\sigma^2.$$

Note that

$$\lambda_\pm = -ik\Omega \pm \sigma_1.$$

7.3.2 A Few Conclusions

We have seen (Section 7.1 and figures therein) that the compliance is smallest at the supports, which means that $S(x)$ is largest there; and we showed, for $\varepsilon = 0$, that the vertical displacement due to a constant static force is inversely proportional to $S_0^{\frac{1}{2}}$ or ν (see Equation (9)). A reasonable approximation to $S(x)$ in practice is:

$$S(x) = S_0(1 + \varepsilon\cos kx) \qquad (18)$$

so that $f(x) = -\cos kx$. In this case the Fourier coefficients of f are all zero except for $a_1 = \bar{a}_{-1} = -\frac{1}{2}$. The solution is then given by

$$y_2(\bar{x},\bar{t}) = \zeta_1(\bar{x},\bar{t}) + \overline{\zeta_1(\bar{x},\bar{t})}$$

$$= 2\,\mathrm{Re}[\zeta_1(\bar{x},\bar{t})]$$

$$= \frac{-P_0\sigma}{2Tk\,d}\left\{e^{-\sigma|x|}[k\cos k(\bar{x} + v\bar{t}) \pm 2\sigma\sin k(\bar{x} + v\bar{t})]\right.$$

$$\left. + \frac{\sigma}{\sigma_1} e^{-\sigma_1|x|}[k(1 + 2\Omega)\cos k(v\bar{t} - \Omega\bar{x}) \mp 2\sigma_1 \sin k(v\bar{t} - \Omega\bar{x})]\right\};$$

and the full solution $\bar{y} = y_1 + \varepsilon y_2$ becomes

$$\bar{y}(\bar{x},\bar{t}) = \frac{P_0\sigma e^{-\sigma|x|}}{2T\nu^2} - \frac{\varepsilon P_0\sigma}{2Tk\,d}\left\{e^{-\sigma|x|}[k\cos k(\bar{x} + v\bar{t}) \pm 2\sigma\sin k(\bar{x} + v\bar{t})]\right.$$

$$\left. + \frac{\sigma}{\sigma_1} e^{-\sigma_1|x|}[k(1 + 2\Omega)\cos k(v\bar{t} - \Omega\bar{x}) \mp 2\sigma_1\sin k(v\bar{t} - \Omega\bar{x})]\right\}. \qquad (19)$$

One of the quantities of interest is the vertical displacement of the pantograph at time t. Ideally we require this to be constant, for the reasons outlined in the Introduction. The required quantity is obtained simply from Equation (19) by putting $\bar{x} = 0$, and is

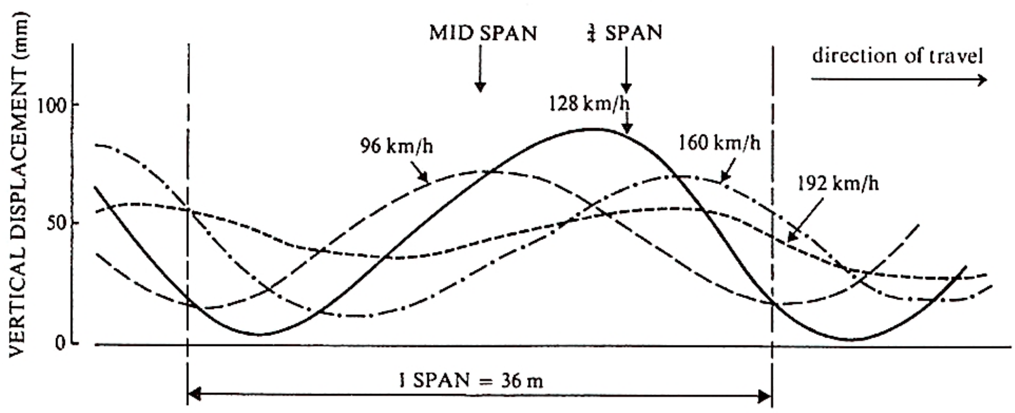
$$\frac{P_0\sigma}{2T\nu^2} - \frac{P_0\sigma\,\varepsilon}{2T\;d}\left[1 + \frac{\sigma(1 + 2\Omega)}{\sigma_1}\right]\cos kvt. \qquad (20)$$

This gives the vertical displacement of the pantograph as it moves along the wire. The displacement as a function of the distance $x(= vt)$ from a given support is

$$\frac{P_0\sigma}{2Tv^2}\left\{1 - \frac{v^2\varepsilon}{d}\left[1 + \frac{\sigma(1 + 2\Omega)}{\sigma_1}\right]\cos kx\right\}. \tag{21}$$

We notice that since $\sigma_1 = 0$ when $v = c(1 + k^2/v^2)^{-\frac{1}{2}}$ the displacement becomes infinite at this velocity. This is clearly wrong, and is a consequence of neglecting the damping term: this error was predicted in Section 7.1. Apart from this the theory predicts (Equation (21)) that the maximum displacement is at the centre of the span (at $x = \pi/k = l/2$) for all velocities.

In practice this is not true and to make better predictions a more sophisticated model is needed in which the motion of the pantograph is taken into account. The results of better calculations are shown in the figure for a reduced scale model.



7.3.3 Multiple Pantographs

Equation (19) gives the disturbance of the wire as seen from the train. To an observer stationary with respect to the wire, the disturbance is given by using the original coordinates,

$$x = \bar{x} + v\bar{t}, \qquad t = \bar{t},$$

where x is now measured along the wire and the train is at $x = vt$.

This gives

$$y(x, t) = \frac{P_0\sigma}{2Tv^2}e^{-\sigma|x-vt|} - \frac{\varepsilon P_0\sigma}{2Tkd}\left\{e^{-\sigma|x-vt|}(k\cos kx \pm 2\sigma\sin kx) + \frac{\sigma}{\sigma_1}e^{-\sigma_1|x-vt|}\{k(1 + 2\Omega)\cos[k(vt(1 + \Omega) - \Omega x)] \mp 2\sigma_1\sin[k(vt(1 + \Omega) - \Omega x)]\}\right\}. \tag{22}$$

Since $\sigma_1 < \sigma$ the dominant oscillations of the wire, for high speeds at least, are given by the last term. When v is large it can be seen that a train with two or more pantographs could excite violent motion of the wire. To see how this can happen, we fix our attention on one point of the wire; the wire at this point moves up and down with a period

$$\frac{2\pi}{kv(1 + \Omega)} = \frac{l}{v(1 + \Omega)}.$$

Now suppose that l_p is the distance between pantographs; then the time between the passing of each pantograph will be l_p/v. If this be equal to an odd number of half periods the resulting disturbance will be smaller; but if it be equal to an integer

number of periods it will reinforce the motion. The condition for this latter is

$$\frac{l_p}{v} = \frac{nl}{v(1+\Omega)} \qquad n \text{ an integer,}$$

or

$$l_p = nl\left(1 - \frac{v^2}{c^2}\right).$$

Because of this dependence on velocity it is not possible to arrange the pantographs to be well behaved at all speeds.

7.4 SUMMARY

A simple mathematical model of the motion of an overhead electric wire has been obtained. The model is very crude in the sense that the point supports have been approximated by a continuous support, and the pantograph has been represented by a constant vertical force. These two approximations are severe, but they enable analytic solutions to be obtained from which a general qualitative idea of the motion may be obtained. In fact the model presented here has been simplified further in order to reduce the algebra to a minimum, but the essential features are still retained. The original paper by Gilbert and Davies takes account of more than we do; in particular some of the dynamical properties of the pantograph are included.

More sophisticated mathematical models have been made but computers are necessary for their solution. These models, and associated experiments, have helped in the understanding of the wire and pantograph motion which has enabled costs of construction and maintenance to be reduced considerably.

7.5 SOLUTIONS TO SELF-ASSESSMENT QUESTIONS

Solution to SAQ 1

(a) The equation of the wire in this instance takes the form

$$\frac{\partial^2 y}{\partial x^2} - \nu^2 y = \frac{g}{c^2},$$

which can be written as

$$\frac{\partial^2}{\partial x^2}\left(y + \frac{g}{\nu^2 c^2}\right) - \nu^2\left(y + \frac{g}{\nu^2 c^2}\right) = 0.$$

The boundary conditions are $y(0, t) = y(l, t) = 0$. The general solution to this (ordinary) differential equation is

$$y(x, t) = -\frac{g}{\nu^2 c^2} + A \sinh \nu x + B \cosh \nu x,$$

and substituting in the boundary conditions we obtain

$$A = \frac{g}{\nu^2 c^2}\left(\frac{1 - \cosh \nu l}{\sinh \nu l}\right), \quad B = \frac{g}{\nu^2 c^2}.$$

From these it follows that

$$y(x, t) = \frac{g}{\nu^2 c^2}\left\{\frac{\sinh \nu(l - x) + \sinh \nu x - \sinh \nu l}{\sinh \nu l}\right\},$$

where we have used the identity

$$\sinh \nu(l - x) = \sinh \nu l \cosh \nu x - \cosh \nu l \sinh \nu x.$$

(b) From the symmetry of the problem it is clear that the maximum sag is midway between the supports, but this may be shown directly. At the point of maximum sag we have

$$\frac{\partial y}{\partial x}(x, t) = \frac{g}{\nu c^2}\left[\frac{-\cosh \nu(l - x) + \cosh \nu x}{\sinh \nu l}\right]$$
$$= 0.$$

Hence

$$\cosh \nu x = \cosh \nu(l - x),$$

which has a root at $x = l/2$. At this point

$$y(l/2, t) = -\frac{g}{\nu^2 c^2}\left[1 - \frac{2 \sinh(\nu l/2)}{\sinh \nu l}\right].$$

(c) From the table in the text we find that

$$\nu l = 26.66, \text{ and } \quad \frac{g}{\nu^2 c^2} = 0.55 \text{ cm}.$$

Now, for $x \gg 1$,

$$\sinh x \simeq \tfrac{1}{2}e^x$$

(e.g. when $x = 4$, $\sinh x = 27.29$, $\frac{1}{2}e^x = 27.30$).

So, approximately,

$$2\frac{\sinh(\nu l/2)}{\sinh \nu l} \simeq 2\frac{e^{\nu l/2}}{e^{\nu l}} = 2e^{-\nu l/2} \simeq 2 \times 10^{-6}.$$

Thus at the mid point the displacement is approximately

$$-\frac{g}{\nu^2 c^2} = -0.55 \text{ cm}.$$

Solution to SAQ 2

(a) A function g is periodic with period l if

$$g(x+l) = g(x).$$

We know that S is periodic, so that

$$S_0(1 - \varepsilon f(x+l)) = S_0(1 - \varepsilon f(x))$$

from which

$$f(x+l) = f(x).$$

Consequently the function f is periodic with period l.

(b) A natural way of representing such a function is by a Fourier series (see *Unit 6*)

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2n\pi x}{l} + b_n \sin \frac{2n\pi x}{l} \right)$$

where

$$a_n = \frac{2}{l} \int_0^l f(x) \cos \frac{2n\pi x}{l} \, dx \qquad n = 0, 1, 2, \ldots,$$

$$b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{2n\pi x}{l} \, dx \qquad n = 1, 2, \ldots.$$

(c) Since the mean value of $f(x)$ is zero,

$$\int_0^l f(x) \, dx = 0,$$

so that $a_0 = 0$ and

$$f(x) \sim \sum_{n=1}^{\infty} \left(a_n \cos \frac{2n\pi x}{l} + b_n \sin \frac{2n\pi x}{l} \right).$$

Solution to SAQ 3

When $\varepsilon = 0$, $S(x) = S_0 =$ constant, and the elasticity of the supporting medium is independent of x. This means that the wire is supported in the same way at every point. This is clearly a bad approximation for the "trolley" wire, but a better approximation for the compound system (see figures in Introduction).

Solution to SAQ 4

If $y(x, t) = e^{i(kx - \omega t)}$ we have

$$\frac{\partial^2 y}{\partial x^2} = -k^2 y, \qquad \frac{\partial^2 y}{\partial t^2} = -\omega^2 y.$$

Substitution into Equation (4) gives

$$\left(k^2 - \frac{\omega^2}{c^2} + v^2 \right) y = 0$$

and, since y is not the zero function,

$$k = \left(\frac{\omega^2}{c^2} - v^2 \right)^{\frac{1}{2}}.$$

The velocity is given by

$$\frac{\omega}{k} = \frac{c}{\left(1 - \frac{v^2 c^2}{\omega^2} \right)^{\frac{1}{2}}}.$$

Solution to SAQ 5

The equation

$$-\frac{\partial^2 z}{\partial x^2} + \frac{1}{c^2}\frac{\partial^2 z}{\partial t^2} + v^2 z = 0$$

admits complex solutions of the form

$$z(x, t) = e^{-i\omega t}(Ae^{ikx} + Be^{-ikx}),$$

where $$k = \left(\frac{\omega^2}{c^2} - v^2\right)^{\frac{1}{2}}.$$

At the origin the above solution has the form

$$z(0, t) = (A + B)e^{-i\omega t} = e^{-i\Omega t},$$

so that $A + B = 1$, $\omega = \Omega$.

If $\Omega > cv$, k is real and since $z(x, t) = z(-x, t)$, $A = B$, and the solution is

$$z(x, t) = \cos kx\, e^{-i\Omega t},$$

the real part of which is $\cos kx \cos \Omega t$.

If $\Omega = cv$, then $k = 0$ and the solution is simply

$$y(x, t) = \cos \Omega t \text{ for all } x,$$

so that the wire moves up and down bodily.

If $\Omega < cv$, k is imaginary; and writing $k = iK$ where $K = (v^2 - \Omega^2/c^2)^{\frac{1}{2}}$ the general solution is

$$\omega = e^{-i\Omega t}(Ae^{-Kx} + Be^{Kx}).$$

Since the amplitude of the motion is bounded we must have $A_- = 0$ for $x < 0$ and $B_+ = 0$ for $x > 0$, so that the condition at the origin gives $B_- = 1$ and $A_+ = 1$ respectively. The real part of the solution is then

$$y(x, t) = e^{-K|x|}\cos \Omega t,$$

so that the motion dies out rapidly as we move away from the origin.

Solution to SAQ 6

To change variables we use the chain rule for a function of several variables.

$$\frac{\partial y}{\partial x} = \frac{\partial \bar{y}}{\partial \bar{x}}\frac{\partial \bar{x}}{\partial x} + \frac{\partial \bar{y}}{\partial \bar{t}}\frac{\partial \bar{t}}{\partial x},$$

$$\frac{\partial y}{\partial t} = \frac{\partial \bar{y}}{\partial \bar{x}}\frac{\partial \bar{x}}{\partial t} + \frac{\partial \bar{y}}{\partial \bar{t}}\frac{\partial \bar{t}}{\partial t},$$

where $\bar{y}(\bar{x}, \bar{t}) = y(x, t)$. Since

$$\bar{x} = \frac{x - ut}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}, \qquad \bar{t} = \frac{t - \frac{u}{c^2}x}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}},$$

we have

$$\frac{\partial \bar{x}}{\partial x} = \frac{1}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}, \qquad \frac{\partial \bar{x}}{\partial t} = \frac{-u}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}},$$

and

$$\frac{\partial \bar{t}}{\partial t} = \frac{1}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}, \qquad \frac{\partial \bar{t}}{\partial x} = \frac{-\frac{u}{c^2}}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}.$$

Hence,

$$\frac{\partial y}{\partial x} = \frac{1}{\left(1-\dfrac{u^2}{c^2}\right)^{\frac{1}{2}}}\left(\frac{\partial \bar{y}}{\partial \bar{x}} - \frac{u}{c^2}\frac{\partial \bar{y}}{\partial \bar{t}}\right)$$

and

$$\frac{\partial y}{\partial t} = \frac{1}{\left(1-\dfrac{u^2}{c^2}\right)^{\frac{1}{2}}}\left(-u\frac{\partial \bar{y}}{\partial \bar{x}} + \frac{\partial \bar{y}}{\partial \bar{t}}\right).$$

Repeated application of the operators $\partial/\partial x$ and $\partial/\partial t$ gives

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{\left(1-\dfrac{u^2}{c^2}\right)}\left(\frac{\partial^2 \bar{y}}{\partial \bar{x}^2} - \frac{2u}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{x}\,\partial \bar{t}} + \frac{u^2}{c^4}\frac{\partial^2 \bar{y}}{\partial \bar{t}^2}\right)$$

$$\frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} = \frac{1}{\left(1-\dfrac{u^2}{c^2}\right)}\left(\frac{u^2}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{x}^2} - \frac{2u}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{x}\,\partial \bar{t}} + \frac{1}{c^2} + \frac{\partial \bar{y}}{\partial \bar{t}}\right).$$

and subtraction gives

$$\frac{\partial^2 \bar{y}}{\partial \bar{x}^2} - \frac{1}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{t}^2} = \frac{\partial^2 y}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} = 0.$$

Solution to SAQ 7

(a) The general solution to the wave equation can be written in the form

$$y(x, t) = F(ct - x) + G(ct + x).$$

In this case, because there is a point force acting at the origin, there may be different solutions in the domains on either side of the origin.

Consider first $x > 0$. The two parts of the above solution represent two distinct motions: $F(ct - x)$ represents a disturbance travelling away from the origin and $G(ct + x)$ a disturbance travelling towards the origin. In this problem the only disturbance affecting the wire is at the origin, and physically we know that any disturbance must travel away from here: consequently the solution is

$$y(x, t) = F(ct - x) \qquad x > 0.$$

The boundary condition at the origin gives,

$$y(0, t) = f(t) = F(ct).$$

Thus

$$y(x, t) = f\left(\frac{ct - x}{c}\right) \qquad x > 0.$$

A similar reasoning gives

$$y(x, t) = f\left(\frac{ct + x}{c}\right) \qquad x < 0.$$

We note that the displacement of the wire is symmetric about the origin, i.e.

$$y(-x, t) = y(x, t) \qquad x, t \in R.$$

(b) We see immediately that

$$\frac{\partial y}{\partial x}(-x, t) = -\frac{\partial y}{\partial x}(x, t) \qquad x, t \in R$$

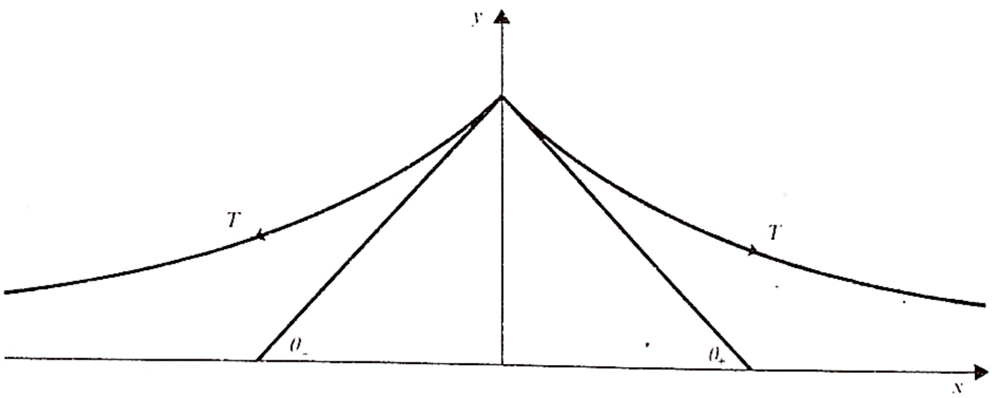
Let $-\tan\theta_+$, $\tan\theta_-$ be the gradients at the origin, as shown. Then

$$\tan\theta_+ = -\left(\frac{\partial y}{\partial x}\right)_{0^+},$$

$$\tan \theta_- = \left(\frac{\partial y}{\partial x}\right)_0 .$$

and so

$$\theta_+ = \theta_- .$$



Thus the horizontal component of the force at $x = 0$ vanishes. The vertical component is given by

$$-T(\sin \theta_+ + \sin \theta_-) = -2T \sin \theta_- .$$

Since the wave equation is only valid when $\partial y/\partial x$ is small we can write

$$\sin \theta_- \simeq \tan \theta_- = \left(\frac{\partial y}{\partial x}\right)_0$$

$$= \frac{1}{c} f'(t).$$

The vertical force is then

$$-\frac{2Tf'(t)}{c}.$$

Solution to SAQ 8

Let x be the distance along the wire travelled by the ring at time t. The imposed condition is given at the position occupied by the ring, and we therefore seek a coordinate system $(\bar{x}, \bar{t})$ in which this position becomes $\bar{x} = 0$. The transformation of SAQ 6 is such that when $x = ut$, $\bar{x} = 0$. In the transformed coordinate system, the imposed condition $y(ut, t) = a \sin \omega t$ becomes

$$\bar{y}(0, \bar{t}) = a \sin \left[\frac{\omega \bar{t}}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}} \right] = g(\bar{t}), \text{ say,}$$

where $\bar{y}(\bar{x}, \bar{t}) = y(x, t)$ is the transverse displacement of the wire at the point given by

$$\bar{x} = \frac{x - ut}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}, \qquad \bar{t} = \frac{t - \frac{u}{c^2} x}{\left(1 - \frac{u^2}{c^2}\right)^{\frac{1}{2}}}.$$

The motion of the wire in (x, t) coordinates is governed by the wave equation. Hence, by SAQ 6, $\bar{y}(\bar{x}, \bar{t})$ satisfies the same equation. The motion subject to a disturbance $g(\bar{t})$ at the origin $\bar{x} = 0$ is, by SAQ 7,

$$\bar{y}(\bar{x}, \bar{t}) = g\left(\frac{c\bar{t} - \bar{x}}{c}\right) \qquad \bar{x} > 0,$$

$$= g\left(\frac{c\bar{t} + \bar{x}}{c}\right) \qquad \bar{x} < 0.$$

It is easily seen that

$$c\bar{t} - \bar{x} = \frac{c\left(t - \dfrac{u}{c^2}x\right) - (x - ut)}{\left(1 - \dfrac{u^2}{c^2}\right)^{\frac{1}{2}}} = \left(\frac{1 + \dfrac{u}{c}}{1 - \dfrac{u}{c}}\right)^{\frac{1}{2}}(ct - x)$$

and

$$c\bar{t} + \bar{x} = \left(\frac{1 - \dfrac{u}{c}}{1 + \dfrac{u}{c}}\right)^{\frac{1}{2}}(ct + x).$$

Thus the solution is

$$y(x, t) = a \sin \frac{\omega(ct - x)}{c - u} \qquad x > ut,$$

$$= a \sin \frac{\omega(ct - x)}{c + u} \qquad x < ut.$$

The solution for $x > ut$ applies in front of the disturbance and we notice that here the frequency of oscillations in the wire is

$$\frac{\omega}{1 - \dfrac{u}{c}}$$

which is greater than ω, whilst behind the disturbance the frequency has decreased. This is an example of the *Döppler effect*.

We notice also that the solution is not valid when $u = c$.

Solution to SAQ 9

The easiest way to make this transformation is to note that it is the same as that in SAQ 6 with $1/c$ replaced by 0 and u replaced by v, so that

$$\frac{\partial^2 y}{\partial x^2} = \frac{\partial^2 \bar{y}}{\partial \bar{x}^2}$$

and

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 \bar{y}}{\partial \bar{x}^2} - 2v \frac{\partial^2 \bar{y}}{\partial \bar{x}\,\partial \bar{t}} + \frac{\partial^2 \bar{y}}{\partial \bar{t}^2}.$$

Thus the equation of motion becomes

$$-\left(1 - \frac{v^2}{c^2}\right)\frac{\partial^2 \bar{y}}{\partial \bar{x}^2}(\bar{x}, \bar{t}) - \frac{2v}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{x}\,\partial \bar{t}}(\bar{x}, \bar{t}) + \frac{1}{c^2}\frac{\partial^2 \bar{y}}{\partial \bar{t}^2}(\bar{x}, \bar{t})$$
$$+ v^2[1 - \varepsilon f(\bar{x} + v\bar{t})]\bar{y}(\bar{x}, \bar{t}) = 0$$

which is Equation (6).

Solution to SAQ 10

(a) Writing $\sigma = v(1 - v^2/c^2)^{-\frac{1}{2}}$ Equation (7) becomes

$$\frac{\partial^2 \bar{y}}{\partial \bar{x}^2} - \sigma^2 \bar{y} = 0 \qquad \bar{x} \neq 0.$$

Two independent solutions to this are $\bar{y}(\bar{x}, \bar{t}) = e^{\pm\sigma\bar{x}}$, and so the general solution is

$$\bar{y}(\bar{x}, \bar{t}) = Ae^{-\sigma\bar{x}} + Be^{\sigma\bar{x}}.$$

Since there is a point force acting at the origin $\bar{x} = 0$, this point is excluded from the domain. Thus we are solving two separate problems, and the same solution need not necessarily be valid both sides of the origin. We therefore label the constants A and B differently in each domain.

(b) There are two conditions imposed upon this solution since it is supposed to represent a physical wire. Firstly, it must be continuous for all $\bar{x}$ (including $\bar{x} = 0$), and secondly, it must be bounded as $|\bar{x}|$ gets large.

(c) The first condition of (b) shows that

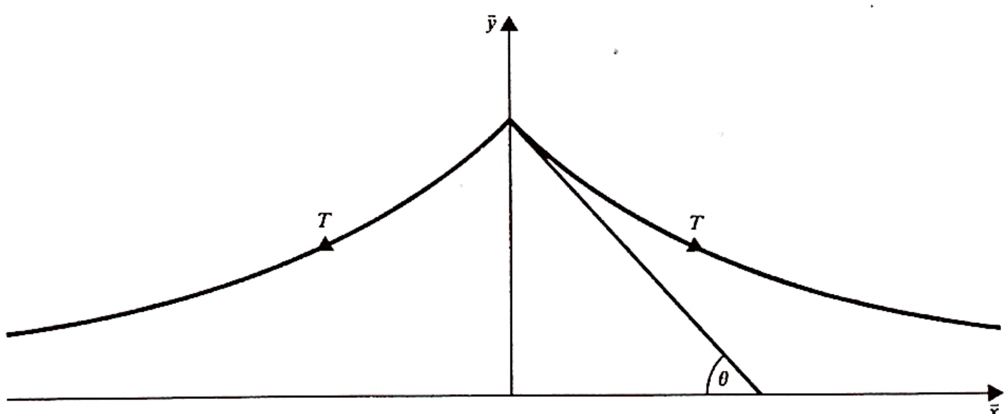
$$A_+ + B_+ = A_- + B_-.$$

The second condition implies that $B_+ = A_- = 0$, so that $A_+ = B_- = A$ (say), and the solution is

$$\bar{y}(\bar{x}, \bar{t}) = Ae^{-\sigma|\bar{x}|}.$$

(d) The constant A can be obtained by resolving the forces at the origin. The upward force is P_0 and the downward force exerted by the tension in the string is $2T\sin\theta$ (note that the wire is symmetric about $\bar{x} = 0$). Now in the derivation of the wave equation, we suppose that $\partial\bar{y}/\partial\bar{x}$ is small, and so

$$\tan\theta = \left(\frac{\partial\bar{y}}{\partial\bar{x}}\right)_{\bar{x}=0^-} \simeq \sin\theta.$$



On equating forces at the origin

$$\begin{aligned} P_0 &= 2T\left(\frac{\partial\bar{y}}{\partial\bar{x}}\right)_{\bar{x}=0^-} \\ &= 2TA\sigma \\ &= 2TA\nu, \end{aligned}$$

since $\sigma = \nu$ when $v = 0$.

Solution to SAQ 11

Putting $\zeta(\bar{x}, \bar{t}) = \exp(\lambda\bar{x} + i\omega\bar{t})$, we obtain

$$\frac{\partial^2\zeta}{\partial\bar{x}^2} = \lambda^2\zeta,$$

$$\frac{\partial^2\zeta}{\partial\bar{x}\,\partial\bar{t}} = i\omega\lambda\zeta$$

and

$$\frac{\partial^2\zeta}{\partial\bar{t}^2} = -\omega^2\zeta.$$

Since $L[\zeta] = 0$ we get a quadratic equation for λ,

$$\left(1 - \frac{v^2}{c^2}\right)\lambda^2 + \frac{2i\omega v}{c^2}\lambda + \frac{\omega^2}{c^2} - \nu^2 = 0.$$

This has solutions λ_+ and λ_- given by

$$\left(1 - \frac{v^2}{c^2}\right)\lambda_\pm = -\frac{i\omega v}{c^2} \pm \nu\left[\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{\nu^2 c^2}\right]^{\frac{1}{2}}.$$

If

$$\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{v^2 c^2} > 0,$$

the real parts of $\lambda_\pm$ are positive and negative respectively. Since the solution is bounded throughout the whole domain it must take the form

$$\begin{aligned} \zeta(\bar{x}, \bar{t}) &= A \exp(\lambda_- \bar{x} + i\omega \bar{t}) \qquad \bar{x} > 0, \\ &= A \exp(\lambda_+ \bar{x} + i\omega \bar{t}) \qquad \bar{x} < 0, \end{aligned}$$

representing damped wave propagation in either direction.

If, on the other hand,

$$\left(1 - \frac{v^2}{c^2}\right) - \frac{\omega^2}{v^2 c^2} < 0,$$

both λ_+ and λ_- are imaginary and the solution is of the form

$$\zeta(\bar{x}, \bar{t}) = A \exp i(\beta \bar{x} + \omega \bar{t}) \qquad \beta, \omega \in R,$$

the real and imaginary parts of which represent undamped waves.

PARTIAL DIFFERENTIAL EQUATIONS OF APPLIED MATHEMATICS

1 *W* The Wave Equation

2 *W* Classification and Characteristics

3 *W* Elliptic and Parabolic Equations

4 NO TEXT

5 *S* Finite-Difference Methods I: Initial Value Problems

6 *W* Fourier Series

7 *N* Motion of Overhead Electric Train Wires

8 *S* Finite-Difference Methods II: Stability

9 *W* Green's Functions I: Ordinary Differential Equations

10 *W* Green's Functions II: Partial Differential Equations

11 *S* Finite-Difference Methods III: Boundary Value Problems

12 NO TEXT

13 *W* Sturm-Liouville Theory

14 *W* Besel Functions

15 *N* Finite-Difference Methods IV: Further Topics

16 *N* Blood Flow in Arteries.

The letter after the unit number indicates the relevant set book: N indicates a unit not based on either book.

Course Team

Chairman:	Professor R. C. Smith	Professor of Mathematics
Members:	Dr. A. Crilly	B.B.C.
	Mr. D. W. Jordan	University of Keele
	Dr. A. D. Lunn	Lecturer in Mathematics
	Dr. N. P. Mett	Lecturer in Mathematics
	Dr. A. G. Moss	Lecturer in Educational Technology
	Dr. D. Richards	Lecturer in Mathematics
	Mr. M. G. T. Simpson	Course Assistant
	Dr. P. Smith	University of Keele
	Dr. P. G. Thomas	Lecturer in Mathematics
	Dr. R. V. Zahar	Senior Lecturer in Mathematics

With assistance from:

	Mr. P. Dewar	Staff Tutor in Mathematics
	Professor L. Fox	Oxford University
	Dr. M. W. Green	University of Dundee
	Professor A. Jeffrey	University of Newcastle-upon-Tyne
	Mr. J. E. Phythian	Staff Tutor in Mathematics
	Mr. G. D. Smith	Brunel University
	Dr. T. B. Smith	Lecturer in Physics
	Mr. G. Young	Staff Tutor in Mathematics

0 335 01251 5

